STUDIA INFORMATICA

Formerly: *Zeszyty Naukowe Politechniki Śląskiej, seria INFORMATYKA*
Quarterly

Volume 32, Number 1B (95)

Krzysztof A. CYRAN

ARTIFICIAL INTELLIGENCE, BRANCHING PROCESSES AND COALESCENT METHODS IN EVOLUTION OF HUMANS AND EARLY LIFE

Silesian University of Technology Press
Gliwice 2011

STUDIA INFORMATICA **Volume 32, Number 1B (95)**

Formerly: *Zeszyty Naukowe Politechniki Śląskiej, seria INFORMATYKA* Nr kol. 1841

Editor in Chief

Editorial Board

STUDIA INFORMATICA is indexed in INSPEC/IEE (London, United Kingdom)

PL ISSN 0208-7286, QUARTERLY
Printed in Poland
The paper version is the original version

If we knew what are we looking for,

It would not be called research, would it?

Albert Einstein

I WOULD LIKE TO DEDICATE THIS BOOK

TO MY WIFE, MY CHILDREN AND MY PARENTS,

WITHOUT WHOM I COULD NOT BE MYSELF

AKNOWLEDGEMENTS

Several persons and organizations have contributed to creation of this book and the author wishes to express his gratitude. First is Professor Marek Kimmel from William Marsh Rice University, Houston, USA who has helped the author in studying the exciting world of bioinformatics and evolutionary genetics by supervising his post-doc visit at Department of Statistics at Rice. The next is Professor Adam Mrózek, who, before his premature death, has introduced the author to the theory of rough sets and its applications. The author also would like to thank the reviewers of this monograph, Professor James F. Peters, Professor Stanisław Cebrat, and Professor Andrzej Polański, for their comments and suggestions, which helped to avoid some errors and make the final version more readable. The list of others would be long as is the list of author's collaborators at different stages of his research, including co-authors of scientific papers, author's supervisors, and reviewers from all over the world as well as his colleagues from the Institute of Informatics at the Silesian University of Technology, Gliwice, Poland. Since contemporary research requires funds, the author would also like to thank the funding institutions, especially those who financed his scientific projects and habilitation grant. In particular, the author would like to acknowledge the fact that this part of the scientific work described in the book, which was performed during last two years, was financed by Polish Ministry of Science and Higher Education from funds for supporting science in 2008-2010 as a research project number N N519 31 9035.

CONTENTS

SPIS TREŚCI

MATHEMATICAL NOTATIONS

A	a set
$a \in A$	element of a set
$\{a_1, a_2, \ldots, a_n\}$	a set consisting of elements $a_1, a_2, \ldots, a_n$
$x_1, x_2, \ldots, x_n$	independent variables
U	set of universe
$\varnothing$	empty set
$\subset, \subseteq, \supset, \supseteq$	relation of containment for sets
$\cap, \cup$	intersection and union for sets
$\wedge, \vee$	conjunction and disjunction for statements
$\otimes$	Exclusive-OR operator
$\times$	Cartesian product operator
$\neg$	negation for statements and set elements
$\{x: *\}$	a set of points satisfying condition $*$
f, g, F	function (general symbol)
$F(x)$	function of variable x
$G \circ F$	superposition of mappings (functions)
$\rightarrow$	assignment, functional dependence
$\overset{k}{\rightarrow}$	dependence at the k^{th} level
0, 1	identity elements in Boolean algebra
$=$	equality relation
$\equiv$	identity relation
$<, \leq, >, \geq$	less than (or equal), greater than (or equal) relations
$\approx$	approximate equality relation
$\neq$	inequality relation

$\Leftrightarrow$, $\equiv$, iff	equivalence (if and only if)
$\Rightarrow$	implication
$\forall$	for all
$\exists$	there exists
R	relation (general symbol)
$x\,R\,y$	x is in relation R with y
$I(Q)$	indiscernibility relation with respect to set of attributes Q
$[x]_{I(Q)}$	abstract class of the relation $I(Q)$ containing element x
$\underline{Q}X$	lower approximation of a set
$\overline{Q}X$	upper approximation of a set
$\overline{w}$	average value of w
$\hat{p}$	estimate of a variable p
$P^x(Y)$	prob. of Y when starting branching process from x elements
$\sim$	asymptotic equivalence
$card\,(X)$	cardinality of a set X
$RED\,(C)$	set of all reducts of a set C
$RED_R\,(C)$	set of all relative reducts of a set C
$RED^x\,(C)$	set of all value reducts of a set C
$RED_R{}^x\,(C)$	set of all relative value reducts of a set C
$CORE\,(C)$	core of the set of attributes C
$CORE_R\,(C)$	relative core of the set of attributes C
$CORE^x\,(C)$	value core of the set of attributes C
$CORE_R{}^x\,(C)$	relative value core of the set of attributes C
$\nabla^2 u$	Laplacian of the function u
$\nabla^2 \mathbf{G}$	vector Laplace operator of a vector field $\boldsymbol{G}$
$\|\| x\|\|$	norm of x
■	end of proof
▬	end of definition

ACRONYMS AND ABBREVIATIONS

A	Adenine
ADALINE	Adaptive Linear Elements
AfAm	African American
AI	Artificial Intelligence
ANA	Alanyl Nucleic Acids
ANN	Artificial Neural Network
ASPM	Abnormal Spindle-like Microcephaly-associated
ATM	Ataxia Telangiectasia Mutated
B	Wall's neutrality test *B*
BASC	BRCA1-associated genome surveillance complex
BF	Binary Fission distribution
BLAST	Basic Local Alignment Search Tool
BLM	Bloom Syndrome
blm^{Ash}	Mutation in BLM
BP	Branching Process
BRCA1	Breast Cancer 1 gene
C	Cytosine
cDNA	Complementary DNA
CGH	Computer Generated Hologram
CI	Computational Intelligence
CM	Coalescent Model
CRSA	Classical Rough Set Approach
*D**	Fu and Li's neutrality test *D**
DOVD	Diffractive Optical Variable Device
DRSA	Dominance-based Rough Set Approach
EA	Evolutionary Algorithm
EM	Expectation-Maximization
*F**	Fu and Li's neutrality test *F**
FOXP2	speech-related gene FOXP2
F_s	Fu's neutrality test F_s
FS	Fuzzy Sets
G	Guanine
GC	Granular Computing
GKP	Granular Knowledge Processing
GNA	Glycol Nucleic Acids
HKA	Hudson-Kreitman-Aguade's neutrality test
H. Neanderthalensis	Homo Neanderthalensis
hRPA	Human Replication Protein A
HRWD	Holographic Ring Wedge Detector

H. Sapiens	Homo Sapiens
IAM	Infinite Allele Model
ISM	Infinite Sites Model
KDE	Kernel Density Estimator
LF	Linear Fractional distribution
LVQ	Learning Vector Quantization
MADALINE	Multiple Adaptive Linear Elements
MDTOG	Maximal amount of Different Types Of Genes
MLP	Multi Layer Perceptron
MNH	Multi-Null-Hypotheses
MRCA	Most Recent Common Ancestor
mtDNA	Mitochondrial DNA
mtEve	Mitochondrial Eve
NORM	Number of Replicating Molecules
NS	Non Significant
NST	Near Set Theory
P	Poisson distribution
PCR	Polymerase Chain Reaction
PDF	Probability Density Function
PGF	Probability Generating Function
PNA	Peptide Nucleic Acid
PNN	Probabilistic Neural Network
p-RNA	Pyranosyl Analog of Ribose
Q	Wall's neutrality test *Q*
QDRSA	Quasi Dominance-based Rough Set Approach
RBF	Radial Basis Function
RECQL	RECQL helicase gene
RMS	Root Mean Square
RS	Rough Sets
RST	Rough Set Theory
RWD	Ring Wedge Detector
RUG	Random Union of Gametes
RUZ	Random Union of Zygotes
S	Strobeck's neutrality test *S*
SCS	Soft Competition Scheme
SIPF	Salt-Induced Peptide Formation
SNP	Single Nucleotide Polymorphism
SOM	Self-Organizing Map
SSMM	Symmetric Stepwise Mutation Model
T	Tajima neutrality test
T	Thymine
TNA	Threose Nucleotide Analogs
U	Uracil
VPRSA	Variable Precision Rough Set Approach
VQ	Vector Quantization
W-F	Wright-Fisher
WRN	Werner Syndrome
WTA	Winner Takes All
WTM	Winner Takes Most
Z_{nS}	Kelly's neutrality test Z_{nS}

1. INTRODUCTION

1.1. Problem genesis

In the post-genomic era the huge amount of genetic data obtained from the Human genome project, Common Chimpanzee genome project, Neanderthal genome project, as well as the currently started 1000 Genomes project, requires development of new advanced methods and technologies for processing and understanding these data. This is an important challenge for information sciences and it motivates both, the form and the content, of this book. In particular, the book is focused on artificial intelligence (AI) and computer simulations whose applicability have already been proven to be of importance for evolutionary genetics. In this context, three research domains have been described: (a) development of artificial intelligence and computer simulations methods used for detection of natural selection at molecular level, (b) stochastic models for estimation of genetic interactions between *H. sapiens* and *H. Neanderthalensis*, including mitochondrial Eve controversy, and (c) computer simulation models of the early stages of the RNA-world.

The book will therefore deal with the earliest and the latest stages of biological evolution: the origin of life, and the evolution of humans. However, the contribution to information sciences inspired by author's research projects is not limited to these particular applications. Rather, the methods presented are tested against these real and biologically sound problems with a clear potential to benefit applications in a much wider and general context of information sciences.

The current state-of-the-art in one of the most rapidly developing artificial intelligence branches, called computational intelligence (CI), is characterized by an enormous progress in the fields of artificial neural networks (ANN), evolutionary algorithms (EA), as well as fuzzy sets (FS) and granular computing (GC). One of the prominent theories in GC is the rough set (RS) theory founded by Pawlak (1982, 1992) which is a basis for development of other approaches such as variable precision rough sets approach (VPRSA) proposed by Ziarko (1993), dominance-based rough sets approach (DRSA) proposed by Greco, Matarazzi and Slowinski (1999a), or near sets model (NSM) proposed by Peters (2007). These

generalizations and modifications constitute the state-of-the-art within granular knowledge processing (GKP).

In this context, the book will present an original approach developed by the author (Cyran 2009d), called quasi-dominance rough set approach (QDRSA). Similarly, the current state-of-the-art in stochastic model simulations, characterized by a wide use of the Monte Carlo methodology, is a background for the software developed and used by the author for efficient simulation of branching processes (BP) in forward time. Challenges for information sciences involved in such simulations are discussed further on subsequent pages of the monograph.

1.2. Organization of the dissertation

The whole book is composed of two parts, the first, dedicated for presenting the methods, and the second, focused on an application of the methods described in part one to the real, biologically sound problems of evolutionary genetics. Part one contains two chapters: chapter 2, devoted to artificial intelligence, and chapter 3, describing the coalescent method and branching processes theory using a background of population genetic models. Part two is composed of three chapters: chapter 4, focused on the neutral theory of evolution with emphasized problem of the search for signatures of natural selection, chapter 5, presenting a human evolution, in particular an application of branching processes methods in the genealogy of mitochondrial DNA (mtDNA) polymorphism of modern humans and their interactions with Neandertals, and chapter 6, discussing the origins of Life with special attention devoted to the information content in the RNA-world hypothetical proto-species. Finally, chapter 7 serves as a summary, which presents the overall conclusions, draws plans for further directions of the research, and speculates about possible results.

The above description of the structure of the book is supplied with the information below, organized in a less formal way. In particular, the order of the chapters will not be treated as a criterion for order of presented issues. Rather, the problems which are tackled in the book are given in their wide context, and appropriate fragments of the book which deal with these problems are identified. Both descriptions of the content, structural, and problem-related (the later detailed also in section 1.4), complement each other and serve as a two-way guide for the reader.

The problem-focused description of the book starts with explanation of the relevance of natural selection studies. It is well known that the proper treatment of complex genetic disorders requires reliable results from association studies, and thus the effective screening for candidate genes exhibiting signatures of natural selection at molecular level. Such screening methods, as presented in chapter 4 of the book, can be based on mutations in genes implicated in human familial cancers caused by instability of DNA replication. The search for

an effective screening procedure for genes under pressure of natural selection constitutes a relevant socio-economic reason for such and similar research. The developed AI-based screening technologies will add-up to the more reliable and time effective search for human genes shaped by natural selection, as targets for possible association with complex genetic diseases.

For the scientific community not less important is discovering trajectories of human evolution and simulating the early life models. These studies constitute a clear and biologically sound motivation for chapters 5 and 6 of the book. The author expresses his hope that the methods presented, both, original and reviewed, will contribute proportionally to the limited size of the book to the scientific understanding of such fundamental issues as how life originated and how hominid lineages led to *H. sapiens*.

The AI-based methods, given in chapter 2, are expected to be of importance for the field of artificial intelligence and, in particular, computational intelligence. The rationale is that AI methods developed during author's research projects, while related to evolutionary genetics, have a potential for knowledge acquisition and processing in a much wider spectrum of problems. The progress in AI caused by development of the author's novel QDRSA is expected to go beyond genetic applications, although this approach was tested on the biologically inspired problem.

1.3. Objectives of the dissertation

The reader should take in mind that the book has been written by a computer scientist and therefore it has been done from an information processing perspective. However, not surprisingly, the multidisciplinary aspects of the book are visible, too. In particular, the title of the book, by enumerating artificial intelligence, branching processes and coalescent methods, refers to (1) information sciences, (2) applied probability with a lot of references to algorithmics of computer simulations, and (3) population genetics. The second part of the title indicates the evolution as the area where these methods are applied. The first region of the evolution considered in the book is the origin of humans, the second is the origin of life. Together, they form two problems situated among the most fundamental in the contemporary biology, which raise serious implications for perceiving Nature. Certainly, theories trying to explain them scientifically have to be multidisciplinary. Among others, they must rely on the development of computer science techniques, since, without improving the knowledge processing methods, the extremely large amount of genetic data will lack its explanation and possible verification in simulation studies.

While the current theories concerning the origin of life, despite many important discoveries, are still at a very hypothetical and speculative stage, the studies focused on the

evolution of humans support scientists with the increasingly precise description, based on the experimental evidence of the hominisation process, which led to the appearance of *H. sapiens*. Despite this clear difference in the current status of these two fields, there is a common need for supporting paleontology, biochemistry and genetics with the increasingly effective information processing tools. **This is where advances in information sciences can support not only scientists but also the society at large, especially in the context of the healthcare. Therefore, the objective of the dissertation is the description of methods which the author has developed and/or used in his scientific work related to mentioned above problems of evolutionary genetics**. To keep the form of a monograph, which describes fields of artificial intelligence, branching processes, and coalescent methods applied in evolutionary genetics, the efforts of other scientists in these areas are also reported as a background material. In this aspect, the monograph can be treated as a concise review of the field with emphasized elements which are relevant for the research work carried out by the author.

1.4. Statement of the problems

To be able to describe the three research domains (a), (b), and (c), defined in section 1.1, the appropriate methodological approaches had to be employed by the author in the related research work. Advantages and disadvantages of the novel methods and techniques developed within this work (or still being under development) are summarized in what follows:

a) **Development of methods used for the search of natural selection at molecular level**. The two different methodologies used by the author include multi-null-hypotheses (MNH) method described in section 4.3.2 and AI-based technologies given in section 4.3.3. The advantage of the MNH method is the potential for more accurate inference using statistical testing against null hypotheses with incorporated nonselective effects (population growth, substructure, and recombination), as compared to testing against classical nulls, where nonselective factors often confound the results. The disadvantage is the requirement for intensive computer simulations in order to estimate the critical values for neutrality statistics tested against modified nulls. However, this drawback is an inspiration for applying AI methodology, which eliminates the need for computer simulations. Therefore, the AI-based strategy can be used in a fast screening procedure for the candidate genes, possibly associated with complex genetic diseases. The rule-based and connectionists techniques will be considered as the AI-based methods applied for this goal. Chapter 2, dedicated to artificial intelligence methods, presents both these techniques. In particular, the author's novel concept, quasi-dominance rough set approach

(QDRSA), which is still under development, is presented in section 2.3.3. It is then compared, in section 4.3.3, on the basis of a real, genetic application, with both, DRSA and the classical rough set approach (CRSA). The first author's studies (Cyran 2009d) indicated that QDRSA exhibits advantages for some classes of problems over both, CRSA and DRSA, however more systematic research is required. Within the connectionist techniques, reviewed in section 2.2.1, such as multilayer perceptrons (MLP), Hopfield networks, Kohonen self organizing maps (SOM) and probabilistic neural networks (PNN), this latter approach was considered in the search for natural selection (section 4.3.3). The overall comparison of the rule-based and the connectionist approaches, applied in the search for the best screening technology, will be given in sections 4.3.3 and 4.4 to the extent possible at the current stage of the research.

b) **Development of branching process models for estimating mitochondrial Eve epoch and the limits of Neanderthal mtDNA admixture in the gene pool of the Upper Palaeolithic *H. sapiens***. The effect of genetic drift, which could eliminate the hypothetical mtDNA contribution of Neandertal mtDNA, is modeled by the slightly supercritical Markov's branching process (BP) using the O'Connell model. The theory of branching processes used for discovering gene genealogies, is described in section 3.6. The novelty and the advantage of this methodology lies in the potential for more accurate modeling of the history of Neanderthal mtDNA genes in *H. sapiens* gene pool as compared with models based on the Wright-Fisher (W-F) models with constant population size. Therefore, it is expected to yield more accurate estimates as compared to the existing model proposed by Serre et al. (2004) studying coexistence of *H. sapiens* and *H. neanderthalensis* in Europe 30 000 years ago. The BP-based model can be applied using recent author's development of methods dating the root of mtDNA polymorphism in contemporary humans. Using the results of these methods, which indicate fast convergence to the O'Connell's limits (see section 3.6 and section 5.3), it is possible to reliably estimate the time of Neandertals extinction relative to the time of the most recent common ancestor (MRCA) of mtDNA of modern humans. However, it requires intensive computer simulations for modeling the Markov BPs in forward time. Such simulations constitute a serious algorithmic challenge because of inherent instability of BPs, which either tend to extinction or grow-up to huge population sizes. Nevertheless, the forward-time simulations deserve an increased interest, since not all genetically feasible phenomena can be modeled using the classical backward-time approach, known as the **coalescent method** (described in chapter 3, section 3.5). The advantage of the latter approach is that it eliminates the computational effort required for processing and storage of all extinct lineages. In the O'Connell model the notion of coalescence is reformulated in terms of BP genealogy. Moreover, with the increase of computer power, both in terms

of the speed and of the memory size, the forward-time simulations, being able to encompass evolution of more and more generations, gain constantly growing interest in the real, genetically inspired, problems such as these considered in the book. Relevance of this particular research lies in treating mtDNA-based studies as complementary approaches to those based on nuclear DNA sequenced in the Neandertal genome project. This project produced the first results in 2006 (Green et al. 2006) and recently, a draft Neandertal genome was sequenced within it (Green et al. 2010).

c) **Development of the models of early stages of the RNA-world**. The methodology is based on the intensive computer simulations of several models, including the compartment model with random segregation of the genetic material. The early life models are given in chapter 6, and the compartment model in section 6.3 of this chapter. The improvement to the existing approaches lies in the modeling of the environmental changes, which affect the evolving population by stochastic fluctuation of the number of replicating molecules (NORM) in the compartment. This stochasticity can be the sole source of variation or it can be added to the cell-to-cell stochasticity originally proposed by Niesert (1987). Further enhancement relying on BP extinction conditions applied to simulated population of RNA protocells is also possible, but it is still under the developmental stage. The relation of this study to the book's content lies in the development of computer simulation algorithms with random number generators requiring extremely large range of aperiodicity. The aim is to model the evolution of the early RNA-world before the appearance of the chromosomal architecture of genomes. Additionally, the conditions of the transition from abiotic to biotic world are considered.

Finally, the comparison of the single-strand models (described in sections 6.2 and 6.4) and the compartment model (described in section 6.3) is carried out in section 6.5 from the information processing perspective, by using the Shannon information theory. The potential of models for preserving the genetic information is studied for the compartment and the single strand models with the complexity threshold estimated in Demetrius-Kimmel BP model supplemented by the author with parameter denoting the probability of the phosphodiester bond break. The advantage of this latter model lies in its potential for obtaining reliable estimates of its parameters. Since the probability of the break of a phosphodiester bond between two nucleotides can be experimentally received for feasible conditions of the early Earth, the model can be more accurate than models based on information balance between mutation and natural selection. Advantageous in the proposed comparison is also the use of *information amount* as a measure of evolutionary capacity of hypothetical models of the RNA-world.

The efficient research in the multidisciplinary studies, such as these covered in this book, demands skills in computer science, probability and statistics, and genetics – therefore there is always a risk that some of these fields will not be treated appropriately. However, this risk has to be taken for all problems located at the interface between information sciences and genetics, the two technological and scientific disciplines that drive a significant part of contemporary innovation. It is a challenge for contemporary scientists, and in particular for the author, to work with those methodologically different disciplines and this book is personal and definitely subjective response to this challenge.

PART I

METHODS

2. ARTIFICIAL INTELLIGENCE

2.1. Foundations

Intelligent machines have occurred in human imagination for hundreds of years, however it is only since the last century, when this imagination has given the birth of a scientific area called artificial intelligence (AI). This is a branch of computer science, probably as old as the computer science itself – the model of artificial neuron, proposed by McCulloch and Pitts (1943) or a formulation of the Turing (1950) test of intelligence can be considered as the beginning of the field, although the name *artificial intelligence* has been introduced a few years later by McCarthy who organized in 1956 the Dartmouth Summer Research Conference on Artificial Intelligence.

During more than 50 years of a development of the field, the philosophy of AI has formulated three fundamental questions (see Russell and Norvig 2003). The first, which is the most important for computer science, is whether a machine with sufficient computational power and large enough memory is able, after appropriate programming, to act intelligently in a sense that it can solve any problem which can be solved by a thinking human. The second, more philosophical, is the question whether a machine can have a mind and consciousness, in particular a self awareness, and can it feel in a way similar to humans. The positive answer to this question can bring serious ethical issues, summarized in the third question, as to what extent a thinking machine will deserve a special treatment.

While today the third question is a domain of science-fiction writers, the constant development in computational power and memory capacities will support the hardware platform for answering the second mentioned question in a few, or perhaps several, decades on an experimental ground (Kurzweil 2005). These philosophical questions have got the consequences also for cognitive scientists, who try to answer if human brain is essentially a computer – certainly different from that proposed by von Neumann, definitely much complex than that proposed by connectionists, but in principle nothing more than a computer of a still unknown architecture and information processing paradigm.

The above problems leave space for speculations and hypotheses, which can be summarized in two views referred to as a strong artificial intelligence and a weak artificial intelligence. These views are characterized by Russell and Norvig (2003) in the following words: "The assertion that machines could possibly act intelligently (or, perhaps better, act as if they were intelligent) is called the *weak AI* hypothesis by philosophers, and the assertion that machines that do so are actually thinking (as opposed to simulating thinking) is called the *strong AI* hypothesis."

In other words, the strong AI hypothesis assumes that a machine, which is a physical symbol system can have a mind, consciousness and mental states (Searle 1999). Searle distinguished this position from what he called weak AI, and what is summarized in a statement that: “A physical symbol system can act intelligently”. The strong version of AI will be considered in the last chapter of the book – all other chapters while referring to AI, will do so in the meaning of a weak AI form.

A distinction is usually made between the kind of high level symbols that directly correspond with objects in the world, and the more complex "symbols" that are present in an artificial neural network. Early AI research, currently referred to as *good old fashioned artificial intelligence* (GOFAI) was focused on high level symbols. However, there is a number of arguments against symbol processing, which show that human thinking does not consist, or at least it does not consist solely, of high level symbol manipulation. In principle, these arguments do not deny the possibility of strong artificial intelligence, but rather they state that for achieving that stage more than symbol processing is required.

One important argument comes from Gödel (1931) who has proved that it is always possible to create statements which could not by proved neither disproved by a formal system (such as an AI program). Penrose (1989) expanded on this argument speculating that quantum mechanical processes inside individual neurons gave humans special advantage over purely symbolic machines. This will be discussed further in chapter 7. However, Russell and Norvig (2003) point out that Gödel's theorem only applies to what can be proved theoretically, given an infinite amount of memory and time. In practice, all machines (including humans treated as machines) have always finite resources and therefore they have difficulties with proving many theorems which in principle can be proven. Yet, it is not necessary to be able to prove everything in order to have the intelligence.

The second type of argument against symbolic AI is given by Dreyfus ([31]) who noted that human intelligence and expertise depends also on unconscious instincts and not only on conscious symbolic manipulation. He argued that these unconscious skills would never be able to be implemented in formal rules. Turing (1950) argued, anticipating the response to Dreyfus argument, that, just because we don't know the rules that govern a complex behavior, this does not mean that no such rules exist. Later, Russell and Norvig (2003) noted that, in the

years since Dreyfus published his critique, progress has been made towards discovering the "rules" that govern unconscious reasoning.

They indicated that, contrary to GOFAI, the computational intelligence (CI) paradigms, such as artificial neural networks (ANN), evolutionary algorithms (EA) and others, are mostly directed at simulated unconscious reasoning and learning. Therefore, AI research in general has moved away from high level symbol manipulation of GOFAI, towards new models intended to capture more of unconscious reasoning or dealing with uncertainty inherently present in many non trivial human inferences.

In contemporary CI field, several models are explored. They belong to connectionism represented by artificial neural networks, computationalism represented by fuzzy sets (FS) and rough sets (RS) approaches, and population-based models with evolutionary computation (EC) and swarm intelligence (SI). Some of these approaches can be joined, what gives the emergence of neural-fuzzy or evolutionary-fuzzy systems (Łęski 2008).

Out of this spectrum, only those methods which were used in the research work of the author will be described in more detail. They all belong to the CI and they are perceived by the author as representatives of either biologically inspired AI or methods based on formal logic, such as the rule-based AI. The composition of Chapter 2 is influenced by this natural discrimination between these categories. Methods inspired by biology, which are represented by connectionism of neural networks and population-based processing of evolutionary computing, are described in section 2.2. Methods based on formal logic, such as rule-based information systems represented by various rough set models are given in section 2.3.

Certainly, it is author's full responsibility that out of many currently studied machine learning methods, he has subjectively chosen in his research neural and evolving systems as those which had arisen from contemplation of life and the rough set theory as the formal logic-based method. However, after this choice has been done and reflected in his studies, the composition of Chapter 2 could not be different. That is also an explanation why the last section in this chapter is a case study – its goal is to illustrate how in one practical application, all these three approaches have found their place.

More specifically, in the mentioned case study presented in section 2.4, the modified by the author indiscernibility relation is used in a hybrid, opto-electronic recognizer of the Fraunhofer diffraction patterns. The study presents how artificial neural networks can interplay with formal logic of rough sets and with population-based optimization using evolutionary computation. Moreover, this application presents the potential of author's modification of indiscernibility relation described in section 2.3.2. With some exceptions, the modification can find many more applications, especially, that it can be equally well adopted in a generalized, variable precision rough set model (VPRSM), introduced by Ziarko (1993), to meet requirements of analysis of huge data sets. In the application described in section 2.4,

the modified rough sets are used in the evolutionary optimization of the optical feature extractor implemented as a holographic ring-wedge detector. The classification of feature vectors is performed by a probabilistic neural network (PNN), described in section 2.2.1.

2.2. Biologically inspired artificial intelligence methods

The Life, which occurred on the Earth some 3.5 billion years ago (see chapter 6) is the example of the enormously complex information processing system. Therefore, it is not a surprise that many systems which can be observed in the living organisms became the inspiration for researchers working in information sciences. In particular, two (out of many) methods, which are classified as computational intelligence, are described in the following two sections. These are artificial neural networks and evolutionary computation.

Before presenting the details, the author wants to express his reservation about the use of a word *intelligence* in this context. This word is well established in the field (see section 2.1), and that is the reason why the author uses it as a technical term of weak AI. However, because this word is also often overused in many not scientific texts claiming to be scientific, or to have at least scientific background, it is worth to stress that *intelligence*, as a technical term of a weak AI approach, has rather loose connection to what it means in philosophy or in a strong AI – and this is the strong AI, which is omnipresent in science-fiction literature.

While this reservation seems to be true for artificial neural networks, it is even more evident in the case of evolutionary computation. The latter is a powerful technique of *adaptation*, but, unless *intelligence* is considered just as *adaptation* as promoted by Fogel et al. (1966) and Fogel (1997a), one can hardly find anything what resembles *intelligence* in the evolutionary process (except, maybe, the intelligence of the programmer designing the evolutionary world, and the product of biological evolution). Whether the products of artificial evolution can be intelligent in a sense wider than, being simply adaptive, is an open question, and because of enormous development of computational and memory abilities of contemporary computers, it is hoped to be answered soon.

2.2.1. Artificial neural networks

Information processing in natural biological nerve systems has become the inspiration for building artificial structures with similar in some aspects properties, although with the use of simplified elements (Tadeusiewicz 2007). The most complex biological information processor is of course human brain, the only system complex enough for making possible the occurrence of self-consciousness.

Tadeusiewicz (1993) summarizes the brain physical parameters in the context of the processing information speed. Human brain's volume is only 1.4 l., its surface is

approximately 2000 cm^2, and the typical weight is around 1.5 kg. The part of a brain, which is responsible for logical activity is cerebral cortex, having thickness of only 3 mm. Despite such compactness the number of nerve cells in a brain oscillates around 10^{10}-10^{11}, and, what seems to be even more important, the number of connections (synapses) between neurons is between 10^{14} and 10^{15}. The huge number of extremely small information processors (neurons) is in a opposition with a speed of operation of a single neuron. The typical nerve cell impulses have frequency 1-100Hz, duration 1-2 ms, and the voltage 100mV. Therefore, the maximum speed of brain, computed as a number of synapse switching per second, achieves a rate of 10^{15} connections × 100Hz = 10^{17} operations/s. When the processing of sensual perception is considered, the fastest of the senses, the visual channel, operates at a speed 100Mb/s (Tadeusiewicz 1993).

The history of artificial neural networks started with the work of McCullough and Pitts (1943) who proposed the mathematical model of artificial neuron (see Fig. 1), as an element operating according to

$$n_i = \sum_{j=0}^{n} w_{ij} x_j, \quad y_i = \mathbf{1}(n_i) \tag{2.2:1}$$

where n_i is the network excitation, x_j are the inputs for $j = 1, 2, \ldots, n$ and $x_0 = 1$, w_{ij} are weights (corresponding to synapses in biological nerve systems) connecting the receiving neuron i with the source neuron j, y_i is the output of the neuron, and $\mathbf{1}(n)$ is the Heaviside step function, which is a discontinuous function whose value is zero for non-positive argument and one for positive argument. The Heaviside step function, proposed by McCulloch and Pitts to be used in their artificial neuron, is one of possible activation functions, i.e. functions which generate the output of the artificial neuron, based on the value of the network excitation.

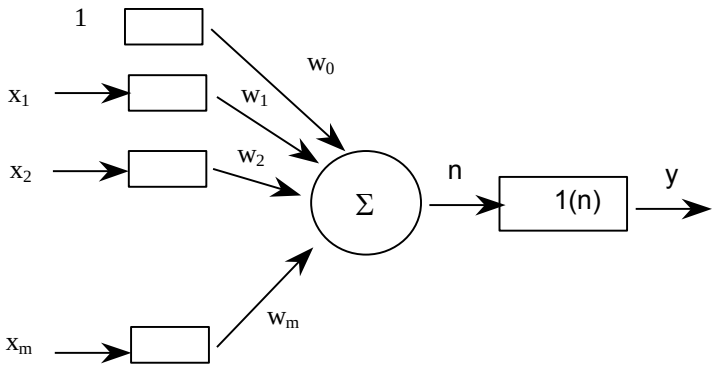


Fig. 2.2:1. McCulloch-Pitts artificial neuron
Rys. 2.2:1. Sztuczny neuron Mc Cullocha-Pittsa

During the history of neural networks other activation functions have been proposed, both, linear and nonlinear, with the sigmoid function, given by (Żurada 1992)

$$y_i = \frac{1}{1+\exp(-\beta n_i)} \tag{2.2:2}$$

where β is a parameter responsible for the slope of the function around network excitation equal zero. The sigmoid function is being most often used due to its non-linearity, differentiability, and continuity. Also for large values of β, it approximates arbitrarily close the Heaviside function.

By grouping artificial neurons with sigmoid activation function in layers, a multiplayer perceptron (MLP) network is obtained, which is the most universal neural network architecture. The neurons in all layers of the MLP are fully interconnected with neurons of the next layer. The connections correspond to synapses in nerve systems, and they are implemented as vectos of weights. The input layer does not process any information, it serves only as a buffer. The last layer produces outputs which are considered as outputs of the whole MLP. Between input and output layer, the arbitrary number of hidden layers can occur, although it is known (see for example Osowski 1996) that a network with two hidden layers can solve a classification problem in arbitrary complex feature space.

A few years after proposition of mathematical model of the first artificial neuron, Hebb (1949) has proposed the coincidence rule for learning such element. Later a lot of different learning rules have been developed, both, for supervised, and unsupervised learning. They all can be described as a product of two functions *g* and *h*, which can be considered as a learning rule, which in general can be dependent on network excitation n_i, desired value on the output d_i, the actual output O_i , and the weight w_{ij} . This general learning rule is given by

$$\Delta w_{ij} = g(n_i, d_i)h(O_j, w_{ij}). \tag{2.2:3}$$

The unsupervised learning rule uses the function *g* in formula (3) which is not dependent on d_i, while the supervised learning rule uses the function *g* which depends on the desired value d_i. For example the unsupervised Hebb's rule given by (Hebb 1949)

$$\Delta w_{ij} = \eta n_i O_j \tag{2.2:4}$$

is a special case of (3) with $g = \eta n_i$, and $h = O_j$. Similarly, Widrow and Hoff (1960) supervised delta rule given by

$$\Delta w_{ij} = \eta(d_i - n_i)O_j \tag{2.2:5}$$

and applied to Adaptive Linear Elements (ADALINE), assumes $g = \eta(d_i - n_i)$, and $h = O_j$.

While ADALINE and Multiple ADALINE (MADALINE) were linear neural networks, the Rosenblatt (1958) proposed a perceptron, which was the nonlinear network. In nowadays classification the Rosenblatt's perceptron is considered as a very reduced version of MLP network, however, it should be mentionded that it was in fact the first neural network ever implemented and it was used for recognition of alphanumerical characters. The perceptron

was built as an electronic – electromechanic system and Rosenblatt has proven that of the solution of the problem exists, then, the perceptron can be trained using the convergent algorithm.

The very fruitful for artificial neural networks two decades have been finished with a Minsky and Papert (1969) famous book, criticizing the connectionist approach as appropriate only for linearly separable problems, and therefore, inappropriate for as simple problems as the exclusive OR function. This critique was addressed to one layer artificial neural networks but it has resulted in a decade of stagnancy of the whole field. The rebirth of interest in ANNs is connected with works showing that nonlinear multilayered networks are free from the limitations signaled by Minsky and Papert for one layered perceptrons. The additional, deciding step toward contemporary artificial neural networks has been done by development of a back-propagation algorithm (Rumelhart, Hinton, and Williams 1986a, 1986b, and Rumelhart et al. 1992) – an efficient method for supervised training of MLP. The derivation of back-propagation algorithm implementing the steepest descent method, is presented below after Tadeusiewicz (1993) and Lawrence (1994).

Let $\{(\mathbf{x}^{(1)}, \mathbf{d}^{(1)}), ..., (\mathbf{x}^{(L)}, \mathbf{d}^{(L)})\}$ be a training set. Observe that superscripts in parentheses denote the number of the training facts for which the learning occurs. The error E computed for the whole training set is a sum of errors for all training examples. It follows that

$$E = \sum_{l=1}^{L} E^{(l)}, \tag{2.2:6}$$

where $E^{(l)}$ is the error of the ANN for the l-th training given by formula

$$E^{(l)} = \sum_{m=1}^{M} E_m^{(l)} = \frac{1}{2} \sum_{m=1}^{M} (d_m^{(l)} - y_m^{(l)})^2, \tag{2.2:7}$$

in which $E_m^{(l)}$ is the error of the m^{th} neuron for the l^{th} training fact.

Definition 2.2:1 (Learning of the neural network)

The learning of the neural network is a minimization of error E in a space of weights w_{ij}.

■

Since, even the simplest networks have a huge number of weights, it is minimization of a scalar field over a space with hundreds (or thousands) of dimensions. To minimize E the steepest descent, gradient-based, method is used.

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} = -\eta \sum_{l=1}^{L} \frac{\partial E^{(l)}}{\partial w_{ij}}. \tag{2.2:8}$$

The above equation indicates that the modification of weights is performed after presenting the whole training set, however often, for algorithm simplicity, the weights are modified after each training fact with appropriately smaller value of the parameter η, called

the learning rate. This parameter should be a positive number, equal typically less than one. To large value of the learning rate can cause the oscillation around the minimum of the error function, too small value results in slow convergence. When modification after each training fact is applied, then (8) should be replaced by an equation, which is indexed by the training fact number *l*. Therefore,

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial E^{(l)}}{\partial w_{ij}}. \tag{2.2:9}$$

Since error generated by network does not directly depend on weights, but on output values, and these values are subsequently dependent on weights, therefore the chain rule is applied

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial E^{(l)}}{\partial w_{ij}} = -\eta \frac{\partial E^{(l)}}{\partial O_i^{(l)}} \frac{\partial O_i^{(l)}}{\partial w_{ij}} = -\eta \frac{\partial E^{(l)}}{\partial O_i^{(l)}} \frac{\partial O_i^{(l)}}{\partial n_i^{(l)}} \frac{\partial n_i^{(l)}}{\partial w_{ij}}. \tag{2.2:10}$$

Using (1) it follows that

$$\frac{\partial n_i^{(l)}}{\partial w_{ij}} = O_j^{(l)} \tag{2.2:11}$$

$$\Delta w_{ij}^{(l)} = -\eta \frac{\partial E^{(l)}}{\partial O_i^{(l)}} \frac{\partial O_i^{(l)}}{\partial n_i^{(l)}} O_j^{(l)}. \tag{2.2:12}$$

Definition 2.2:2 (Generalized delta, after Lawrence 1994)

The generalized delta δ_i of neuron *i* for training example (*l*) is defined as a negative partial derivative of the error $E^{(l)}$ with respect to the network excitation function $n_{(i)}{}^{(l)}$.

■

By applying the chain rule, the generalized delta can be expressed as

$$\delta_i^{(l)} = -\frac{\partial E^{(l)}}{\partial O_i^{(l)}} \frac{\partial O_i^{(l)}}{\partial n_i^{(l)}} \tag{2.2:13}$$

and

$$\Delta w_{ij}^{(l)} = \eta \delta_i^{(l)} O_j^{(l)}. \tag{2.2:14}$$

The meaning of generalized delta depends on the location of the neuron considered. For neurons in output R^{th} layer, denote $O_i^{(l)R} = y_i^{(l)}$. It follows that

$$\begin{aligned} \frac{\partial E^{(l)}}{\partial O_i^{(l)R}} &= \frac{\partial \left\{ \sum_{m=1}^{M} \frac{1}{2} (d_m^{(l)} - y_m^{(l)})^2 \right\}}{\partial y_i^{(l)}} = \frac{1}{2} \sum_{m=1}^{M} \frac{\partial \left\{ (d_m^{(l)} - y_m^{(l)})^2 \right\}}{\partial y_i^{(l)}} = \\ &= \frac{1}{2} \frac{\partial \left\{ (d_i^{(l)} - y_i^{(l)})^2 \right\}}{\partial y_i^{(l)}} = -(d_i^{(l)} - y_i^{(l)}) \end{aligned} \tag{2.2:15}$$

and therefore, for output layer R, after substituting the derivative of the sigmoid (logistic) activation function, the generalized delta is given as

$$\begin{aligned}\delta_i^{(l)R} &= \beta(d_i^{(l)} - y_i^{(l)})O_i^{(l)R}(1-O_i^{(l)R}) = \\ &= \beta(d_i^{(l)} - y_i^{(l)})y_i^{(l)}(1-y_i^{(l)}).\end{aligned} \tag{2.2:16}$$

For hidden layers, the generalized delta has to be computed recursively. Let us start from the last hidden layer with index $R-1$. It follows that

$$\frac{\partial E^{(l)}}{\partial O_i^{(l)R-1}} = \frac{\partial \sum_{k=1}^{M} E_k^{(l)}}{\partial O_i^{(l)R-1}} = \sum_{k=1}^{M} \frac{\partial E_k^{(l)}}{\partial O_i^{(l)R-1}} = \sum_{k=1}^{M} \frac{\partial E_k^{(l)}}{\partial n_k^{(l)R}} \frac{\partial n_k^{(l)R}}{\partial O_i^{(l)R-1}}. \tag{2.2:17}$$

Using (1) it follows also that

$$\frac{\partial n_k^{(l)R}}{\partial O_i^{(l)R-1}} = w_{ki} \tag{2.2:18}$$

and, consequently

$$\begin{aligned}\frac{\partial E^{(l)}}{\partial O_i^{(l)R-1}} &= \sum_{k=1}^{M} \frac{\partial E_k^{(l)}}{\partial n_k^{(l)R}} w_{ki} = \sum_{k=1}^{M} \frac{\partial E_k^{(l)}}{\partial O_k^{(l)R}} \frac{\partial O_k^{(l)R}}{\partial n_k^{(l)R}} w_{ki} = \\ &= \sum_{k=1}^{M} \frac{\partial E^{(l)}}{\partial O_k^{(l)R}} \frac{\partial O_k^{(l)R}}{\partial n_k^{(l)R}} w_{ki} = -\sum_{k=1}^{M} \delta_k^{(l)R} w_{ki}.\end{aligned} \tag{2.2:19}$$

Finally, for the layer $R-1$:

$$\delta_i^{(l)R-1} = \beta O_i^{(l)R-1}(1-O_i^{(l)R-1}) \sum_{k=1}^{M} \delta_k^{(l)R} w_{ki} \tag{2.2:20}$$

and for any layer, the following recursive equation holds

$$\delta_i^{(l)r} = \beta O_i^{(l)r}(1-O_i^{(l)r}) \sum_{k=1}^{N_{r+1}} \delta_k^{(l)r+1} w_{ki}. \tag{2.2:21}$$

Equation (21) uses the back-propagation of generalized deltas in a neural network what is the reason for the name of the whole algorithm. The back-propagation algorithm, described by equations (16), (21), and (14) is universal but slowly convergent error minimization technique. Therefore, this method is often modified by the introduction of the inertial term called momentum (Tadeusiewisz 1993). Then the equation (14) becomes

$$\Delta w_{ij}^{(l)} = \eta \delta_i^{(l)} O_j^{(l)} + \mu \Delta w_{ij}^{(l-1)}, \tag{2.2:22}$$

or, in a version called exponential smoothing (Lawrence 1994),

$$\Delta w_{ij}^{(l)} = \eta((1-\mu)\delta_i^{(l)} O_j^{(l)} + \mu \Delta w_{ij}^{(l-1)}). \tag{2.2:23}$$

The MLP networks trained with back-propagation algorithm with inertial modifications have proved to be one of the most universal networks, applicable for enormous class of practical problems, from pattern recognition, by financial instruments prediction, to medical

diagnosis support. However, these networks were not the only ANNs, which have been developed in the eighties of the twentieth century.

Hopfield (1982) designed the recurrent ANN capable to serve as an autoassociative memory and heuristically solving the traveling salesman problem. This is a network with associated Lapunov energy function (Cohen and Grossberg, 1983) minimized during operation of the network. The structure of the Hopfield network is given Fig. 2. It is noteworthy to mention that the operation of this network can be expressed also in terms of statistical mechanics using the notion of Hamiltonian for denoting the energy function, as shown by Hertz, Krogh, and Palmer (1991).

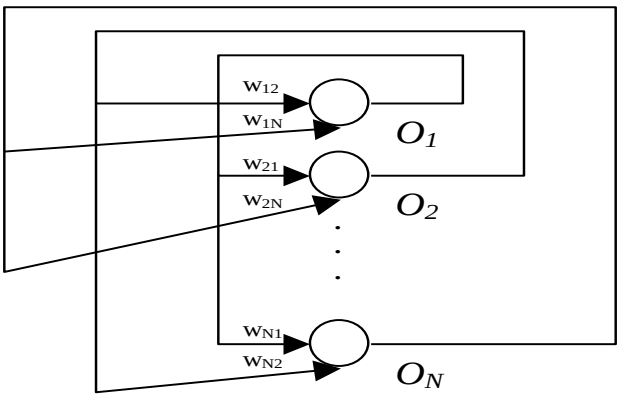


Fig. 2.2:2. Hopfield's network
Rys. 2.2:2. Sieć Hopfielda

The operation of the discrete version of Hopfield's network is described by two formulae as given in Korbicz, Obuchowicz, and Uciński (1994). The first, is the formula used for computation of the network excitation of the neuron k which is randomly chosen for activation

$$n_k^{(p)} = \sum_{j=1}^{N} w_{kj} O_k^{(p)} - t_k \ . \tag{2.2:24}$$

In equation (24) and in all other equations describing the Hopfield's network, the superscripts in parentheses are used to denote the actual step number during the operation of the network, rather than to denote the number of the training fact. The second formula describing operation of Hopfiled's network is used for definition of the activation function. It follows that the output of the network depends on the network excitation as

$$O_k^{(p+1)} = \begin{cases} 1 & \Leftrightarrow n_k^{(p)} > 0 \\ O_k^{(p)} & \Leftrightarrow n_k^{(p)} = 0 \\ 0 & \Leftrightarrow n_k^{(p)} < 0 \end{cases} \ . \tag{2.2:25}$$

Note, that equation (25) defines the activation function in a discrete Hopfield's network, which is used as an autoassociative memory. This activation function is very similar to Heaviside function **1**(*n*), however it takes special interest in the value of the function for $n = 0$. For this situation, the Hopfield's network simply does not change the current output of the neuron, so the new state can be both, 0 or 1, dependent on the present value.

The next two important features, which characterize the Hopfield's network (see Korbicz, Obuchowicz, and Uciński 1994) include the lack of self-dependence

$$\forall i, w_{ii} = 0 \tag{2.2:26}$$

and the symmetry of weights

$$\forall ij, w_{ij} = w_{ji}. \tag{2.2:27}$$

As it has been mentioned, with each Hopfield's network, the so called energy function (Lapunow function) is associated. This is function, which has finite lower bound and which is non-increasing during the evolution of the process considered (in our context, the process of change of states in a recurrent Hopfield's network).

The operation is started for $p = 0$ by connecting the inputs to the processing units. Assuming that the input vector $\mathbf{x} = [x_1, x_2,...,x_N]$, $x_i \in \{0,1\}$, it follows that $O_i(0) = x_i$ for $i = 1, 2, ..., N$. Then the input signals are disconnected and the recurrent operation of the network begins. This process satisfies the equations (24) and (25). The network operates asynchronously, i.e. in a given moment each neuron can be chosen with equal probability, and only this neuron is activated. After a finite number of iterations, the network settles in a stable state, for which

$$\forall i, O_i^{p+1} = O_i^p. \tag{2.2:28}$$

This is a state corresponding to the local minimum of the energy function. This state is transmitted to the outputs of the network.

The energy function is chosen as (Korbicz, Obuchowicz, and Uciński 1994)

$$E(\mathbf{O}) = -\frac{1}{2}\mathbf{O}^T\mathbf{w}\mathbf{O} + \mathbf{t}^T\mathbf{O}. \tag{2.2:29}$$

what, in a scalar notation is equivalent to

$$E(\mathbf{O}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}O_iO_j + \sum_{i=1}^{N} t_iO_i. \tag{2.2:30}$$

Lemma 2.2:1

The energy $E(\mathbf{O})$ is a non-increasing function of time during the operation of a network.

Proof

Let in a moment *p*, the state of the k^{th} neuron be randomly chosen to be changed

$$O_k^{(p+1)} = O_k^{(p)} + \Delta O_k^{(p)}. \tag{2.2:31}$$

Moreover, let the state of others neurons remain unchanged

$$\forall j \neq k, O_j^{(p+1)} = O_j^{(p)}. \tag{2.2:32}$$

Then, it follows that

$$\begin{aligned}\Delta E^{(p)} &= E\left(\mathbf{O}^{(p+1)}\right) - E\left(\mathbf{O}^{(p)}\right) = \\ &= -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} O_i^{(p+1)} O_j^{(p+1)} + \sum_{i=1}^{N} t_i O_i^{(p+1)} + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij} O_i^{(p)} O_j^{(p)} - \sum_{i=1}^{N} t_i O_i^{(p)}.\end{aligned} \tag{2.2:33}$$

Expanding the summations for terms with $j = k$ and $j \neq k$ results in

$$\begin{aligned}&\Delta E^{(p)} = \\ &= -\frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1, j\neq k}^{N} w_{ij} O_i^{(p+1)} O_j^{(p+1)} + w_{ik} O_i^{(p+1)} O_k^{(p+1)}\right) + \sum_{i=1, i\neq k}^{N} t_i O_i^{(p+1)} + t_k O_k^{(p+1)} + \\ &\quad + \frac{1}{2}\sum_{i=1}^{N}\left(\sum_{j=1, j\neq k}^{N} w_{ij} O_i^{(p)} O_j^{(p)} + w_{ik} O_i^{(p)} O_k^{(p)}\right) - \sum_{i=1, i\neq k}^{N} t_i O_i^{(p)} - t_k O_k^{(p)},\end{aligned} \tag{2.2:34}$$

which, after expanding similarly the external summations and canceling corresponding terms for states of neurons different than k^{th}, whose outputs are identical at moments p and $p + 1$, and using the lack of self-dependence (26), becomes

$$\begin{aligned}\Delta E^{(p)} &= -\frac{1}{2}\sum_{i=1, i\neq k}^{N} w_{ik} O_i^{(p+1)} O_k^{(p+1)} - \frac{1}{2}\sum_{j=1, j\neq k}^{N} w_{kj} O_k^{(p+1)} O_j^{(p+1)} + t_k O_k^{(p+1)} + \\ &\quad + \frac{1}{2}\sum_{i=1, i\neq k}^{N} w_{ik} O_i^{(p)} O_k^{(p)} + \frac{1}{2}\sum_{j=1, j\neq k}^{N} w_{kj} O_k^{(p)} O_j^{(p)} - t_k O_k^{(p)}.\end{aligned} \tag{2.2:35}$$

Using (31) for terms for the moment $p + 1$, the equation (35) can be transformed to

$$\begin{aligned}\Delta E^{(p)} &= -\frac{1}{2}\sum_{i=1, i\neq k}^{N} w_{ik} O_i^{(p)} O_k^{(p)} - \frac{1}{2}\sum_{i=1, i\neq k}^{N} w_{ik} O_i^{(p)} \Delta O_k^{(p)} \\ &\quad - \frac{1}{2}\sum_{j=1, j\neq k}^{N} w_{kj} O_k^{(p)} O_j^{(p)} - \frac{1}{2}\sum_{j=1, j\neq k}^{N} w_{kj} \Delta O_k^{(p)} O_j^{(p)} + t_k O_k^{(p)} + t_k \Delta O_k^{(p)} + \\ &\quad + \frac{1}{2}\sum_{i=1, i\neq k}^{N} w_{ik} O_i^{(p)} O_k^{(p)} + \frac{1}{2}\sum_{j=1, j\neq k}^{N} w_{kj} O_k^{(p)} O_j^{(p)} - t_k O_k^{(p)}\end{aligned} \tag{2.2:36}$$

Canceling identical terms, and using the symmetry property (27), the above can be simplified to

$$\begin{aligned}\Delta E^{(p)} &= -\sum_{j=1}^{N} w_{kj} \Delta O_k^{(p)} O_j^{(p)} + t_k \Delta O_k^{(p)} == -\Delta O_k^{(p)}\left(\sum_{j=1}^{N} w_{kj} O_j^{(p)} - t_k\right) = \\ &= -\Delta O_k^{(p)} n_k^{(p)}\end{aligned} \tag{2.2:37}$$

If $n_k^{(p)} = 0$ then based on (37) the energy remains constant, i.e it is not increasing. All possible situations for $n_k^{(p)} \neq 0$ and the corresponding changes of the energy $E(\mathbf{O})$ based on (25) and (37) are presented in Table 1.

Table 2.2:1
Possible changes of the energy function in the Hopfield network
(after Korbicz, Obuchowicz, and Uciński 1994)

$O_k^{(p+1)}$	$O_k^{(p)}$	$\Delta O_k^{(p)}$	$n_k^{(p)}$	$\Delta E^{(p)}$
0	0	0	< 0	0
0	1	-1	< 0	< 0
1	0	1	> 0	< 0
1	1	0	> 0	0

Inspection of the last column in Table 1 assures that $\Delta E^{(p)}$ is always zero or negative: $\Delta E^{(p)} \leq 0$. Hence, $E(\mathbf{O}^{(p+1)}) \leq E(\mathbf{O}^{(p)})$ what should have been proved.

■

Theorem 2.2:1 (after Korbicz, Obuchowicz, and Uciński 1994)

The energy $E(\mathbf{O})$ decreases with every change in a state of the network.

Proof

By Lemma 1 it is clear that the energy cannot increase. Therefore, to prove the theorem it is enough to show that each situation when the energy remains constant corresponds to the situation when the network state is not changed. Then, each state change will result in the energy decrease. Consider first the case when $n_k^{(p)} = 0$. Then from (25) it follows that $O_k^{(p+1)} = O_k^{(p)}$ (i.e., outputs are not changed). This is one of the conditions when $\Delta E^{(p)} = 0$. Other situations for $\Delta E^{(p)} = 0$ can be taken from Table 1. It follows that for the unchanged energy $O_k^{(p+1)} = O_k^{(p)} = 0$ or $O_k^{(p+1)} = O_k^{(p)} = 1$, so the outputs are not changed as well. Hence, for all changes of the network state, the energy decreases.

■

Theorem 2.2:2 (after Korbicz, Obuchowicz, and Uciński 1994)

In the discrete Hopfield network, the minimum energy E_{min} is finite and it is achieved in a finite number of steps.

Proof (after Korbicz, Obuchowicz, and Uciński 1994)

From (30) it follows that

$$\left|E(\mathbf{O})\right| \leq \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|w_{ij}\right| + \sum_{i=1}^{N}\left|t_i\right|. \tag{2.2:38}$$

Because of (38) and the discrete domain of the network outputs it follows that the minimum non-zero change of energy is not infinitesimally small

$$\exists c, \min_{\Delta E \neq 0} |\Delta E| = c > 0. \tag{2.2:39}$$

By Theorem 1, the energy function decreases with each change of the network state. Since from (38) it is clear that the energy function has a finite lower bound and from (39) it follows that each change of E is at least as large as c, therefore the process of approaching towards the finite minimum value of E_{min} has to be composed of a finite number of steps.

■

As the last example of an ANN developed in the decade of a great rebirth of connectionism, let us present the problem of self-organization (Kohonen1984) occurring in the Kohonen (1990) Self-Organizing Map (SOM). This is an unsupervised network, in which only the winning neuron and its „neighborhood" is learned. The main application of SOM is the search for regions in the input space which is activated by similar feature values. The goal of the self-organizing learning is such choice of the weights, which minimizes the expected value of the distortion, measured as an error of approximation of the input vector **x** by the weights of the winning neuron (see Osowski 1996)

$$E = \frac{1}{p} \sum_{i=1}^{p} \left\| \mathbf{x}_i - \mathbf{w}_{w(i)} \right\|, \tag{2.2:40}$$

where $w(i)$ is the index of the neuron, which wins for the input vector x_i, and $w_{w(i)}$ is the vector of weights leading to this neuron.

After learning, the network implements a vector quantization (VQ), i.e., the approximation of an arbitrary vector by a pattern vector, which is the closest to the vector considered. This process is equivalent to the quantization of the input space. Since the operation of quantization is a result of the learning process, it is called learning vector quantization. Let us discuss the unsupervised learning of SOM in more detail.

One of the simplest algorithms, which is able to learn the SOM is the algorithm called *winner takes all* (WTA). The name points out the fact that only the winning neuron, i.e., the neuron for which the distance between its weights vector $\mathbf{w}_w$ and the input vector **x** is the smallest, is subject to learn. It is also worth to notice that WTA algorithm used in connectionist approaches, corresponds to the *K-means* algorithm in classical cluster analysis.

Learning is an adaptation tending to changing the weights of the winner in the direction of **x** (see Osowski 1996)

$$\mathbf{w}_w(k+1) = \mathbf{w}_w(k) + \eta(\mathbf{x} - \mathbf{w}_w(k)). \tag{2.2:41}$$

Note, that if the input vectors are normalized, then the minimum distance between vector $\mathbf{w}_w$ and the input vector **x**, corresponds to the maximum of the dot product $\mathbf{w}_w \cdot \mathbf{x}$.

However, learning only one neuron per one training fact, leads to relatively slow convergence, and therefore the modification, called *winner takes most* (WTA) is more often used. In this generalized version of the Kohonen's SOM, there is introduced the

neighborhood of the winner, which is also modified during learning together with the winner. Additionally, it is possible to introduce the modification which takes into consideration that the neurons become tired after learning, and therefore are not activated in the subsequent moments. This modification is inspired by biology, and its goal of it is to favor neurons with smaller initial activation.

Learning of the Kohonen's map using the WTM algorithm follows according to the formula (Osowski 1996)

$$\mathbf{w}_i(k+1) = \mathbf{w}_i(k) + \eta_i G(i,\mathbf{x})(\mathbf{x} - \mathbf{w}_i(k)) \tag{2.2:42}$$

for all neurons i which belong to the neighborhood S_w of the winner. The neighborhood function G defines the influence of the distance from the winner on the modification strength. By defining function G as

$$G(i,\mathbf{x}) = \begin{cases} 1 & \text{for } i = w \\ 0 & \text{for } i \neq w \end{cases} \tag{2.2:43}$$

where w denotes the index of the winner, the classical WTA algorithm, as a special case of WTM, is obtained.

In the classical Kohonen's map, the neighborhood function $G(i, \mathbf{x})$ is of the form (Osowski 1996)

$$G(i,\mathbf{x}) = \begin{cases} 1 & \text{for } \quad d(i,w) \le \lambda \\ 0 & \text{for others} \end{cases} \tag{2.2:44}$$

where $d(i, w)$ denotes the Euclidean distance between weights vectors of the winner w and the neuron i^{th}. Coefficient λ is a radius of the neighborhood. Its value decreases with the learning of the network. The function G given by (44) defines the so called rectangular neighborhood.

Another type of the function G, which is used in the Kohonen's maps defines a Gaussian neighborhood. In this type of the neighborhood the function $G(i, \mathbf{x})$ is given as (Osowski 1996)

$$G(i,\mathbf{x}) = \exp\left(-\frac{d^2(i,w)}{2\lambda^2}\right). \tag{2.2:45}$$

The Gaussian neighborhood results in better self-organization than the rectangular neighborhood, because the strength of the learning is gradually decreased with the increase of the distance.

While, both, the rectangular and the Gaussian neighborhoods are deterministic functions of the distance $d(i, w)$, the stochastic relaxation algorithm (see Osowski 1996) defines the neighborhood, which neurons belong to with probabilities given by the Gibbs distribution

$$P(i)=\frac{\exp\left(-\frac{\left\|\mathbf{x}-\mathbf{w}_i\right\|^2}{T}\right)}{\sum_{j=1}^{n}\exp\left(-\frac{\left\|\mathbf{x}-\mathbf{w}_j\right\|^2}{T}\right)}. \tag{2.2:46}$$

In the above distribution, T is a parameter called the temperature, which has similar role as temperature in a simulated annealing-based optimization.

When the temperature is high at the initial stage of learning, then all neurons belong to the neighborhood with approximately the same probability, what is reflected by the limit

$$\lim_{T\to\infty} P(i)=\frac{1}{N}. \tag{2.2:47}$$

Later, when the temperature is decreasing, the algorithm becomes more and more deterministic, achieving for very small temperatures behavior resembling the WTA algorithm

$$\lim_{T\to 0} P(i)=\begin{cases}1 & \text{for } i=w\\ 0 & \text{for } i\neq w\end{cases}. \tag{2.2:48}$$

The stochastic relaxation defines the random neighborhood of the rectangular type. Therefore, the function G is given by

$$G(i,\mathbf{x})=\begin{cases}1 & \text{for } P(i)\le P\\ 0 & \text{otherwise}\end{cases}, \tag{2.2:49}$$

where P is a random number taken from the uniform distribution with the range (0,1).

The next algorithm considered is the soft competition scheme (SCS). It is a deterministic version of the stochastic relaxation algorithm, which has better effectiveness than the original probabilistic algorithm (see Osowski 1996). Instead of rectangular neighborhood taken with probability P in stochastic relaxation, the SCS uses the Gibbs distribution (46) as the definition of deterministic function G

$$G(i,\mathbf{x})=P(i). \tag{2.2:50}$$

The last algorithm considered in the context of SOM and the neighborhood function G is the neuron gas algorithm (see Osowski 1996), in which all neurons are sorted according to the increasing distance from the vector **x**. Then the function G is given by

$$G(i,\mathbf{x})=\exp\left(-\frac{m(i)}{\lambda}\right), \tag{2.2:51}$$

where $m(i)$ denotes the rank of the neuron i in a sorted sequence, which starts from 1 for the winner, and λ is a decreasing in time parameter, analogous to the radius of the neighborhood in the Kohonen's classical WTM algorithm. If $\lambda = 0$, then, only the winner is modified, and the algorithm becomes the WTA. Otherwise, the algorithm resembles the fuzzy approach, by associating with each neuron a membership function (51) of belonging to the winner

neighborhood. If the quantization error (40) is the criterion, then the following sequence (from the best to the worst) of the self-organizing algorithms is given by Osowski (1996): Neuron gas, SCS, K-means, classical Kohonen's map.

Despite presented above successes of ANNs in many pattern recognition and other machine learning problems, many scientists were not convinced not having the mathematical theory describing the efficiency of ANN-based classifiers. The response to these reservations has been done in nineties of the previous century by proving the following theorem.

Theorem 2.2:3 (after Tebelskis 1995)

Properly trained ANNs are optimal classifiers in pattern recognition problems using statistical uncertainty model, i.e., the output neurons approximate arbitrary closely posterior probabilities of all classes considered.

Proof (after Tebelskis 1995)

Consider an ANN-based classifier learned with many training facts in a form of pairs $(\mathbf{x}, C_j)$ where $\mathbf{x}$ is the input vector and C_j correct abstract class corresponding to that vector. Let index $j = 1, ..., K$ indicate the abstract class and K be the number of classes. Assume that pairs $(\mathbf{x}, C_j)$ have probability distribution $p(\mathbf{x}, C_j)$. Denote also the values occurring at output neurons by $y_k(\mathbf{x})$ with $k = 1, ..., K$. Then, the required outputs of the ANN, denoted by T_{kj} satisfy

$$T_{kj} = \begin{cases} 1 \Leftrightarrow & k = j \\ 0 \Leftrightarrow & k \neq j \end{cases}. \tag{2.2:52}$$

Learning is the minimization of the RMS error functional E, which is a sum of functionals $F_{\mathbf{x}kj}(y_k) = (T_{kj} - y_k(\mathbf{x}))^2$ over all abstract classes j, all output neurons k, and for all input vectors $\mathbf{x}$ proportional to their probability distribution. Hence,

$$E = \int_{\mathbf{x}} \left(\sum_j p(\mathbf{x}, C_j) \sum_k \left(T_{kj} - y_k(\mathbf{x}) \right)^2 \right). \tag{2.2:53}$$

Functional of RMS error E can be written as

$$E = \int_{\mathbf{x}} \left(\sum_k \left(\sum_j p(\mathbf{x}, C_j) \left(T_{kj} - y_k(\mathbf{x}) \right)^2 \right) \right) = \int_{\mathbf{x}} \left(\sum_k E_{\mathbf{x}k} \right) \tag{2.2:54}$$

where

$$E_{\mathbf{x}k} = \sum_j p(\mathbf{x}, C_j) \left(T_{kj} - y_k(\mathbf{x}) \right)^2. \tag{2.2:55}$$

Since $E_{\mathbf{x}k}$ is always positive for any $\mathbf{x}$ and k, therefore minimization of E given by (53) is equivalent to minimization of functional $E_{\mathbf{x}k}$. Separating in $E_{\mathbf{x}k}$ terms for $j = k$ and $j \neq k$ it follows that

$$E_{\mathbf{x}k} = p(\mathbf{x}, C_k)(1 - y_k(\mathbf{x}))^2 + \sum_{j \neq k} p(\mathbf{x}, C_j) y_k^2(\mathbf{x}) =$$
$$= p(\mathbf{x}, C_k)(1 - 2y_k(\mathbf{x}) + y_k^2(\mathbf{x})) + (p(\mathbf{x}) - p(\mathbf{x}, C_k)) y_k^2(\mathbf{x}) = \quad (2.2:56)$$
$$= p(\mathbf{x}, C_k) - 2p(\mathbf{x}, C_k) y_k(\mathbf{x}) + p(\mathbf{x}) y_k^2(\mathbf{x}),$$

and using the fact that $p(\mathbf{x}, C_k) = p(\mathbf{x})p(C_k|\mathbf{x})$,

$$E_{\mathbf{x}k} = p(\mathbf{x})(p(C_k|\mathbf{x}) - 2p(C_k|\mathbf{x}) y_k(\mathbf{x}) + y_k^2(\mathbf{x})) =$$
$$= p(\mathbf{x})p(C_k|\mathbf{x}) + p(\mathbf{x})(-p^2(C_k|\mathbf{x}) + p^2(C_k|\mathbf{x}) - 2p(C_k|\mathbf{x}) y_k(\mathbf{x}) + y_k^2(\mathbf{x})) = \quad (2.2:57)$$
$$= p(\mathbf{x})p(C_k|\mathbf{x})(1 - p(C_k|\mathbf{x})) + p(\mathbf{x})(p(C_k|\mathbf{x}) - y_k(\mathbf{x}))^2.$$

It is clear that functional E_{xk} is minimized when

$$y_k(\mathbf{x}) = p(C_k|\mathbf{x}), \quad (2.2:58)$$

what ends the proof.

■

From (57) it follows that some inherent error E_{xk}^{min} is characteristic even to the properly trained network. This error, which is the result of possible inconsistencies in training data, is

$$E_{\mathbf{x}k}^{\min} = p(\mathbf{x})p(C_k|\mathbf{x})(1 - p(C_k|\mathbf{x})). \quad (2.2:59)$$

While, as it has been shown, all properly trained ANNs can be used as optimal classifiers in probabilistic uncertainty model, there is a special Radial Basis Function (RBF) neural network dedicated for classification of closed regions in the input space. Among different types of RBF neural networks, there is one specifically designed for the statistical pattern recognition. This is a Probabilistic Neural Network (PNN), which is a kernel density estimator (KDE). This special RBF neural network is devoted to the estimation of probability density functions (PDF).

Consider the set $V_j = \{\mathrm{v}(s) \in \Re^N, 1 \le s \le S_j\}$ of feature vectors belonging to class ω_j. Then the kernel estimation of the conditional PDF $p(\mathrm{v} \mid \omega_j)$ is given by (Jutten 1997)

$$\hat{p}(\mathbf{v}|\omega_j) = \frac{1}{S_j} \sum_{\mathbf{v}(s) \in V_j} \frac{1}{h(s, j)^N} K\left(\frac{\mathbf{v} - \mathbf{v}(s)}{h(s, j)}\right). \quad (2.2:60)$$

where $K(\cdot)$ is some kernel function with width $h(s, j)$.

In a majority of applications, the width is fixed and depends only on S_j. PNN, performing the estimation, is a feed-forward network consisting of the input layer, the pattern layer and the summation layer (Raghu and Yegnanrayana 1998). Each neuron of the pattern layer is connected with every neuron of the input layer and the weight vectors of the pattern layer are equal to the feature vectors present in a training set. Contrary to the pattern layer, the summation layer consisting of M neurons, is organized in a such way, that only one output neuron is connected with neurons from any pattern layer pool.

One of the most important features of all ANNs is their ability to generalize the training examples to unknown data. However, there are two phenomena which must be considered in this context. This is a problem of interference and the locality. The interference of the network is a phenomenon which occurs when the learning in one point of the input space results in forgetting examples associated with some other point of this space. Networks, which are less susceptible to the interference are called spatially local. While the interference is defined for pairs of points in the input space, the locality is a property of the whole input domain.

The important problem in that context is how to assure the plasticity of the ANN, so it can learn new facts without forgetting the old ones. The interference is measured by the influence of the learning at point x on mapping implemented by ANN in point $x' \neq x$, as explained formally below.

Consider a mapping given by $y = f(\mathbf{x}, \mathbf{w})$, where $y \in \Re$ is network's output, $\mathbf{x} \in \mathcal{X}$ is the input vector, $\mathbf{w} \in \mathcal{W}$ is the weight vector, and function $f: \mathcal{X} \times \mathcal{W} \to \Re$ is a continuous mapping implemented by ANN with input domain $\mathcal{X} \subset \Re^n$ and weight domain $\mathcal{W} \subset \Re^m$. Then, learning of the ANN is a process of adaptation of weights $\mathbf{w}$, so that ANN approximates required function $y^* = f^*(\mathbf{x})$. Assume that learning algorithm is of a general form $\Delta\mathbf{w} = \alpha\mathbf{H}(\mathbf{x}, \mathbf{w}, e)$, where $\Delta\mathbf{w}$ denotes the change of the weight vector in the time unit, α is a learining rate, $\mathbf{H}$ a direction of weight change, and $e = y - y^*$ an approximation error. Using these notions it is possible to define interference and locality formally.

Definition 2.2:3 (Interface of neural network, after Weaver, Baird, PolyCarpou 1998)

The interference of the network in a point $\mathbf{x}'$ caused by learning in a point $\mathbf{x}$, denoted as $\mathcal{I}_{f,\mathbf{w},\mathbf{H}}(x, x')$ is defined for the unit approximation error as

$$I_{f,\mathbf{w},\mathbf{H}}(\mathbf{x},\mathbf{x}') = \begin{cases} \lim_{\alpha \to 0} \dfrac{f(\mathbf{x}',\mathbf{w}) - f[\mathbf{x}', \mathbf{w} + \alpha\mathbf{H}(\mathbf{x},\mathbf{w},1)]}{f(\mathbf{x},\mathbf{w}) - f[\mathbf{x}, \mathbf{w} + \alpha\mathbf{H}(\mathbf{x},\mathbf{w},1)]} & \text{if the limit exists,} \\ 0 & \text{otherwise.} \end{cases} \tag{2.2:61}$$

■

Definition 2.2:4 (Locality of neural network, after Weaver, Baird, PolyCarpou 1998)

The locality of the network denoted as $\mathcal{L}_{f,\mathbf{w},\mathbf{H},X}$ is defined as a reciprocal of the averaged over the entire input space squared interference of the network.

$$L_{f,\mathbf{w},\mathbf{H},X} = \left[\int_X \int_X I_{f,\mathbf{w},\mathbf{H}}(\mathbf{x},\mathbf{x}')^2 \, d\mathbf{x} d\mathbf{x}' \right]^{-1}. \tag{2.2:62}$$

■

It is worth to notice that, both, the MLP and RBF networks can arbitrarily accurately approximate the continuous functions with arbitrarily large locality, if there is large enough number of neurons (weights). However, too large number of weights decreases generalization ability, and therefore there must be done some trade-off between generalization and locality.

Finally, let us say that in artificial neural networks there is no separation between program and data, and there is no addressable memory. This is completely different from the classical computers with clear separation (at least logical) between program and data, both residing in addressable memory. However, this is not the only property inherited from biological nerve systems. Also error tolerance in hardware implementations of artificial neural networks is to some extent similar to that encountered in real biological nervous nets. This is one more (not decisive of course) argument for the claim that artificial neural networks imitate essential properties of biological nerve systems, despite substantial simplification of artificial neuron as compared to natural one.

Let us consider error tolerance is more detail. One striking thing that can be observed in the context of error tolerance in biological nerve systems, especially in a brain, is time required for functional operation. Note, that these natural information processors are built in such a way so they can tolerate many errors occurring during the lifecycle of an organism. The architecture of a brain allows for errors (relatively often occurring on separate nerves) to be as harmless as possible. This is contrary to classical artificial information processors. Architecture of a computer is designed to make any hardware and system software error as conspicuous as possible.

The explanation is simple. Since computers cannot in general operate properly in the presence of errors in their core elements, each error should be detected because such situation is almost equivalent to the loss of validity of the results. The reliable operation in the immanent presence of errors, characteristic for biological nervous systems, is to some extent propagated to artificial neural networks, however, as it has been said, it should be clear that only hardware implementations (for example, the optical implementations, see Cyran et al. 2001a) of artificial neural networks can really benefit this property.

2.2.2. Evolutionary computing

The genetic algorithms have been developed as systems, which simulate the biological evolution (see section 4.1), however their applications are focused mainly in optimization of problems having little, if any, connections with biology. The scientific theory of artificial evolutionary systems have been founded in the sixties of the twentieth century, when Holland (1967) developed genetic algorithms, Fogel et al. (1966) proposed evolutionary programming, and Schwefel (1965) introduced the idea of evolutionary strategies. Mimicking the natural evolution results in a terminology of evolutionary computing, which uses such

terms as genotypes, phenotypes, chromosomes, alleles, etc. Let us start with definitions of these notions, as formalized by Radcliffe (1997).

Definition 2.2:5 (Search space, alleles)

Let S be continuous or discrete, finite or infinite set of objects, known as search space, and let A_1, A_2, ..., A_n be finite sets of elements $a_{ki} \in A_i$ called alleles.

▬

Definition 2.2:6 (Representation space)

The representation space I is defined as a Cartesian product of sets A_i

$$I = A_1 \times A_2 \times ... \times A_n \tag{2.2:63}$$

▬

Definition 2.2:7 (Decoding function)

The decoding function d is defined as a function which maps vectors from representation space I to search space S

$$d : I \to S . \tag{2.2:64}$$

▬

Definition 2.2:8 (Representation)

The representation of S is defined as the ordered pair (I, d).

▬

Definition 2.2:9 (Chromosome)

Chromosomes are defined as the elements from the representations space I. Alternatively, chromosomes are referred to as genotypes, since simple haploid genetic models with one chromosome per individual are considered.

▬

Definition 2.2:10 (Genes)

Genes at a locus i are defined as the elements x_i of the chromosome $\mathbf{x} \in I$.

▬

Using definitions 5 and 10, it is clear that gene x_i at a locus i, can take one out of possible values (alleles) a_{ki} from a set of alleles A_i.

Definition 2.2:11 (Extended set of alleles)

For each set of alleles A_i, let us define the extended set of alleles A^*_i as

$$A_i^* = A_i \cup \{*\} . \tag{2.2:65}$$

▬

Definition 2.2:12 (Schema)

Each element of the set Ξ, defined as

$$\Xi = A_1^* \times A_2^* \times ... \times A_n^* \tag{2.2:66}$$

is called a schema $\xi = (\xi_1, \xi_2, ..., \xi_n) \in \Xi$, which describes a set of chromosomes with alleles identical with ξ at all positions i, for which $\xi_i \neq *$

$$\xi = \{x \in I : \quad \forall i \in \{1,2,...n\} \ \ (\xi_i = x_i \quad \vee \quad \xi_i = *)\}. \tag{2.2:67}$$

■

Definition 2.2:13 (Defining positions)

All loci i of the schema ξ for which $\xi_i \neq *$, are called the defining positions.

■

Definition 2.2:14 (Order of the schema)

The order $O(\xi)$ of the schema ξ is defined as a number of defining positions.

■

Definition 2.2:15 (Defining length of the schema)

The defining length $\delta(\xi)$ is the distance between the first and the last defining position.

■

Definition 2.2:16 (Fitness function)

Consider an objective function $F: S \to \Re$. Then the fitness function $f: S \to \Re_+$ is defined as such function that satisfies

$$f(x) = \max_I f \Leftrightarrow F(d(x)) = \underset{d(I)}{opt} F \tag{2.2:68}$$

■

Definition 2.2:17 (Pareto-dominance)

In multi-objective optimization, the fitness function is defined as a vector function **f** = ($f_1, f_2, ..., f_n$), where f_i: $I \to \Re_+$. If each of the functions f_i should be minimized then $\forall x,y \in I$ x dominates over y in Pareto sense if and only if $\forall i \in \{1, 2, ..., n\}$: $f_i(x) \leq f_i(y)$ and $\exists j \in \{1, 2, ..., n\}$: $f_j(x) < f_j(y)$.

■

Definition 2.2:18 (Pareto-optimal solution)

The Pareto-optimal solution $x \in I$, is a solution not dominated by any other solution.

■

Definition 2.2:19 (Genetic operations)

The genetic operations: selection *s*, mutation *m*, and recombination *c* are defined by the following functions $s: I^{\lambda} \to I^{\mu}$, $m: I^{\kappa} \to I^{\lambda}$, and $c: I^{\mu} \to I^{\kappa}$, respectively. Note, that these operators are defined for the whole population.

■

After presenting formal definitions of notions present in evolutionary computation, the evolutionary algorithm will be introduced. Denote by μ and λ sizes of the parent and the child population, respectively. Moreover, let $P(t) = (a_1(t),...,a_\mu(t)) \in I^{\mu}$ be a population, $\mathbf{f}(t) \in \Re_+^{\mu}$, a fitness vector for this population, and function *Evaluate* (*t*), the operation used for computation of the fitness in generation *t*. Specify also sets of parameters Θ_s, Θ_m, Θ_c, for genetic operations *s*, *m*, and *c*, respectively, and denote by τ the criterion of the end, which is dependent on the current population $P(t)$ and the set of parameters Θ_τ.

Then, the optimization performed with the use of evolutionary algorithm (EA) can be expressed in pseudo-code as (Bäck, 1997a)

```
Input:    μ, λ, Θs, Θm, Θc, Θτ
Output:   a*∈P* - the best found individual and/or
              P** - the best found population
t ← 0;
P (t) ←Initialize (μ);
f (t) ← Evaluate (P (t), μ);
while (τ (P (t), Θτ) ≠ true) do
       P′(t) ← c (P (t), Θc);
       P″(t) ← m (P′(t, Θm);
       f (t) ← Evaluate (P″(t), μ);
       P (t+1) ← s (P″(t), f (t), μ, Θs);
       t ← t + 1;
od
```

Probabilistic behavior of the evolving population $P(t)$ in AE can be modeled by the stochastic process, which is a homogeneous Markov model. Therefore, the evolutionary process at step $t_k + 1$, does not depend on state of this process before step t_k, if the state of the process at step t_k is known (Rudolf, 1997).

Specific operation of EA is defined by the details of three genetic operators, formally involved in parameters Θ_s, Θ_m, Θ_c. These operators belong to two functionally different classes. Mutation and recombination are operators responsible for generating new solutions, whereas selection is an operator responsible for the choice of the most fitted solutions with probability higher than those, which are less fitted. Selection can be described by a coefficient called the selection pressure and a closely related coefficient called takeover time, as defined below.

Definition 2.2:20 (Selection pressure, after Grefenstette 1997a)

The selection pressure is defined as a rate of increase of the best individual in a population in the absence of mutation and recombination.

▬

Definition 2.2:21 (Takeover time, after Grefenstette 1997a)

The takeover time is defined as the time required for population to be composed of the copies of the best individual only, assuming selection is the only operation, and there is exactly one best individual at the beginning.

▬

It is evident that when the selective pressure increases, the takeover time decreases, and vice versa. The selection is also dependent on the choice of the fitness function, especially in multi-objective optimization, since the fitness value regulates the probability of survival of particular individual during the evolution.

Formally, the fitness function f is described as a superposition of the scaling s, the objective F, and the decoding d, functions $f = s \circ F \circ d$. Therefore, it follows that

$$f : I \xrightarrow{d} S \xrightarrow{F} \Re \xrightarrow{s} \Re_+ \tag{2.2:69}$$

and the fitness function is always maximized due to existence of scaling function s.

In the multi-objective optimization the vector fitness function has to be scalarized. There exist several methods of scalarization Φ, which satisfy the condition that the final fitness of the given solution is not worse than the scalar fitness of all other solutions dominated in the Pareto-sense (Fonseca and Fleming 1997). Since such mappings are not unique, they require the specification of objective preferences. The most commonly used is a scalarization based on the weighted sum, where the preferences are introduced as values of weights w_k. Such approach is given by (see Fonseca and Fleming 1997)

$$\begin{aligned} &\Phi : \Re^n \to R, \\ &\Phi\big(f(x)\big) = \sum_{k=1}^{n} w_k f_k(x). \end{aligned} \tag{2.2:70}$$

Another strategy used for scalarization is applied in a MINI-MAX method (see Fonseca and Fleming 1997)

$$\begin{aligned} &\Phi : \Re^n \to R, \\ &\Phi\big(f(x)\big) = \max_{k=1,\ldots,n} \frac{f_k(x) - g_k}{w_k}, \end{aligned} \tag{2.2:71}$$

where w_k and g_k are parameters responsible for the introduction of preferences.

Yet another method is used in Pareto-scalarization (Goldberg 1989, Fonseca and Fleming 1997), which is defined by recurrent equations

$$\Phi : \Re^n \to \{1,2,...,\mu\},$$
$$\Phi(f(x_i)) = \begin{cases} 1 & \Leftarrow \neg(f(x_j) p < f(x_i)) \quad \forall j \in \{1,2,...,\mu\} \\ \phi & \Leftarrow \neg(f(x_j) p < f(x_i)) \quad \forall j \in \{1,2,...,\mu\} \setminus \{l : \Phi(f(x_l)) < \phi\}, \end{cases} \tag{2.2:72}$$

where, condition $f(x_i)\ p< f(x_j)$ is satisfied when

$$\forall k \in \{1,...,n\} \qquad f_k(\mathbf{x}_j) \le f_k(\mathbf{x}_i) \quad \wedge \quad \exists k \in \{1,...,n\} : \quad f_k(\mathbf{x}_j) < f_k(\mathbf{x}_i). \tag{2.2:73}$$

This scalarization has no possibility to introduce the preferences, however it guarantees that Pareto postulates are automatically fulfilled. Obviously, after scalarization, the multi-objective optimization becomes single-criterion optimization with scalar fitness function.

Let us now consider the influence of three genetic operators on the evolutionary algorithm behavior, starting with the selection. The most natural is the proportional selection, which resembles the selection occurring in biological evolution. The probability distribution of survival is given in this type of selection as (Grefenstette 1997a)

$$p_{proportional}(i) = \frac{f(i)}{\sum_{i=1}^{\mu} f(i)}, \tag{2.2:74}$$

where $f(i)$ denotes the fitness of the I^{th} individual, and μ denotes the size of size of population.

The takeover time in proportional selection is larger than in many other selection types. For the fitness function $f(x) = x^c$, the takeover time $\tau_{proportional}$ is (Golberg and Deb 1991)

$$\tau_{proportional} = \frac{\mu \ln \mu - 1}{c}. \tag{2.2:75}$$

In the tournament selection there is no need for scaling and this selection is especially easy for parallel implementation. However, it should be taken into account that the takeover time in this selection is one of the shortest, i.e., this selection generates very strong selective pressure. It follows, that for the tournament of the size q performed in a population of the size μ, the takeover time $\tau_{tournament}$ is given by Golberg and Deb (1991) as

$$\tau_{tournament} = \frac{1}{\ln q}(\ln \mu + \ln(\ln \mu)). \tag{2.2:76}$$

It is clear that the takeover time is decreasing (and the selective pressure is growing) with the increase of q. Therefore, in tournament selection the user can easily control the values of these important parameters.

Yet another selection is the one based on ranking of individuals in a population. It is possible to consider this type of selection with linear or nonlinear probability distribution of the survival, but in both cases these distributions are based on rankings (0 for the worst, and

$\mu - 1$ for the best) and not on fitness values of particular individuals. Hence, such selection, similarly to the tournament selection, is invariant with respect to the scale and shift of the fitness function. The linear probability distribution of survival is given by (Grefenstette 1997b)

$$p_{linear\_ranking}(i) = \frac{\alpha_{rank} + \frac{rank(i)(\beta_{rank} - \alpha_{rank})}{\mu - 1}}{\mu} \tag{2.2:77}$$

where α_{rank} is a number of children of the worst, and β_{rank} for the best fitted individual.

For the linear distribution the takeover time is approximately (see Goldberg and Deb 1991)

$$\tau \approx \frac{\ln \mu + \ln(\ln \mu)}{\ln 2} \tag{2.2:78}$$

for $\beta_{rank} = 2$, and

$$\tau \approx \frac{2}{\mu - 1} \ln(\mu - 1) \tag{2.2:79}$$

for $1 < \beta_{rank} < 2$.

The nonlinear distributions are often used as geometrical or exponential distributions. Finally, let us consider the Boltzmann selection based on simulated annealing (Mahfoud 1997). The key concept in this selection is a Boltzmann draw, i.e. comparison of individual i and j, in which individual i is a winner with the logistically given probability (Michalewicz 1992)

$$p = \frac{1}{1 + e^{(f_i - f_j)/T}} \tag{2.2:80}$$

where T is a parameter called a temperature, and f_i and f_j are the fitness function values of individuals i and j, respectively. In each of the above models it is possible to introduce a modification called the elitist strategy. It results in propagating of the best individuals with probability one. Elitist models are useful in optimization problems with a goal of finding a global optimum (Sarma and De Jong 1997).

The first studies concerning genetic algorithms supported view that the recombination is a fundamental operator and mutation is less important. With the advent of evolutionary computation with more complex representations, the role of mutation has become more and more evident, and it was stressed that in principle this operator is able operate without a crossing-over (Bäck 1997a). In this context, Cyran et al. (1997) demonstrated that the mutation used without recombination is able to learn ANN in stochastic evolutionary training. In canonical form of genetic algorithm (see below), the mutation operator is defined on binary vectors $\mathbf{a} = (a_1,...,a_l) \in I = \{0,1\}^l$ of the length l. If we denote by p_m the probability

of mutation, than the mutation operator $m:\{0,1\}^l \rightarrow \{0,1\}^l$ generates new vector $\mathbf{a}'=m(\mathbf{a})$ according to

$$a_i' = \begin{cases} a_i & u > p_m \\ 1 - a_i & u \le p_m \end{cases} \tag{2.2:81}$$

where u is an uniform random variable on [0,1] generated for each $i \in \{1,...,l\}$. For more complex representations the mutation operator may be defined by a lot of variants.

For real-valued vectors $\mathbf{x} \in \Re^n$, a new vector $\mathbf{x}' = m(\mathbf{x})$ is produced by mutation m, which is defined most often as (Fogel 1997b)

$$\mathbf{x}' = \mathbf{x} + \mathbf{M} \tag{2.2:82}$$

where $\mathbf{M}$ is a vector of random variables with expected values equal zero, i.e. $E(\mathbf{x}')=\mathbf{x}$. Michalewicz (1992) proposed an non-uniform mutation, changing in time, as described below. Let real-valued chromosome be defined as a vector $\mathbf{x}$, indexed by time t expressed in number of generations

$$\mathbf{x}(t) = < x_1(t),..., x_N(t) > . \tag{2.2:83}$$

Assuming that element $x_i(t)$ is mutated, the result is $\mathbf{x}(t+1) = (x_1(t+1),...,x_i'(t+1),...,x_N(t+1))$, in which $x_i(t)$ is defined as

$$x_i'(t+1) = \begin{cases} x_i(t) + \Delta(t, UB_i - x_i(t)) & \text{for drawn value } 0 \\ x_i(t) - \Delta(t, x_i(t) - LB_i) & \text{for drawn value } 1 \end{cases}. \tag{2.2:84}$$

In the above formula LB_i and UB_i are the lower and upper bounds of the variable $x_i(t)$, and function $\Delta(t,y)$ takes values from a range $[0,y]$, given that probability of $\Delta(t,y)$ being close to zero is increasing with time t. Hence, the initial mutations have relatively large effects (in order to search the whole space), and then the local search is performed. Michalewicz (1992) proposed function $\Delta(t,y)$ defined as

$$\Delta(t, y) = y(1 - r^{(1-\frac{t}{T})b}) \tag{2.2:85}$$

where r is an uniform random variable from [0,1], T is a maximum number of generations, and b is a parameter, which describes the influence of the generation number on the result of the function.

Recombination is in general a binary operator defined on the Cartesian product of the chromosome representation space. It is a mapping r given by (Booker 1997)

$$r : \mathbf{I} \times \mathbf{I} \xrightarrow{m} \mathbf{I} \times \mathbf{I} \qquad r(\mathbf{a}, \mathbf{b}) = (\mathbf{c}, \mathbf{d}), \qquad \mathrm{m} \in \{0,1\}^l \tag{2.2:86}$$

where

$$c_i = \begin{cases} a_i & \Leftarrow m_i = 0 \\ b_i & \Leftarrow m_i = 1 \end{cases} \qquad d_i = \begin{cases} b_i & \Leftarrow m_i = 0 \\ a_i & \Leftarrow m_i = 1 \end{cases} \tag{2.2:87}$$

The mask vector **m** defines the form of the recombination as one-point or multi-point crossing-over. It is also possible to define the uniform recombination in which the number of crossing-over points is not a constant but each point is determined with probability p_x independently for each position on the chromosome. For chromosomes represented in $\Re^n$ it is possible to apply the arithmetic recombination, which does not exchange genes but it is averaging the gene values. In this type of recombination, two parents, x_1 and x_2, are creating one child x' according to (Fogel 1997c)

$$x_i' = \alpha\, x_{1i} + (1-\alpha) x_{2i} \tag{2.2:88}$$

where α is a number from [0,1]. This operator can be generalized to arbitrary many parents, as an *n*-ary operator defined by

$$x_i' = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \ldots + \alpha_k x_{ki}, \qquad \sum_{j=1}^{k} \alpha_k = 1. \tag{2.2:89}$$

After presenting details of the three genetic operators (selection, mutation, and recombination) let us now consider different types of chromosome representation. The classical representation of chromosomes, the binary vectors $\mathbf{a} = (a_1,\ldots,a_l) \in \{0,1\}^l$ constitute the canonical form of genetic algorithms. Genes of such chromosomes take values from binary allele. This representation is especially useful for implementing pseudo-Boolean optimization problems $F: \{0, 1\}^l \to \Re$. However, it is also possible to apply them to optimization of the type $F: S \to \Re$, where S is a search space having different structure as compared to the chromosome representation space $I = \{0,1\}^l$.

One of the most often encountered problems of this class are problems defined as $f: \Re^n \to \Re$, i.e. problems of optimization of continues parameters. These problems require the discretization of the space of continues variables $x_i \in \Re$ onto $[u_i, v_i]$ such that $u_i \le x_i \le v_i$. Then, each such variable can be represented by binary sequence of the length l_x, which is a sub-sequence of the sequence of the length l. For n variables, it follows that $l = nl_x$. The interpretation of the sub-sequence with indices having values from $(i\text{-}1)l_x + 1$ to il_x as a representation of a binary variable x_i is dependent on different decoding functions. For the natural binary code, the decoding function g is a standard binary decoder Γ, which for variable x_i is given as $\Gamma^i: \{0,1\}^l \to [u_i, v_i]$ according to (Bäck 1997b)

$$\Gamma^i(a_1,\ldots,a_l) = u_i + \frac{v_i - u_i}{2^{l_x} - 1}\left(\sum_{j=0}^{l_x-1} a_{il_x - j}\, 2^j\right). \tag{2.2:90}$$

For decoding of the variables expressed in the Gray's code (in which representations of the consecutive numbers are the binary vectors with Hamming distance equal one) the following formula is used (Bäck 1997b)

$$\Gamma^i(a_1,...,a_l) = u_i + \frac{v_i - u_i}{2^{l_x} - 1}\left\lfloor \sum_{j=0}^{l_x-1}\left(\bigotimes_{k=1}^{l_x-j} a_{(i-1)l_x+k} \right) 2^j \right\rfloor. \quad (2.2:91)$$

Such mappings are not able to discern different values of variable x_i if these values lie within the range $\Delta x_i = (v_i - u_i) / (2^{lx} - 1)$. Moreover, since these mapping introduce additional nonlinear transformations in computing the effective objective function F': $\{0,1\}^l \rightarrow \Re$, given as (Bäck 1997b)

$$F'(\mathbf{a}) = (F \circ \bigtimes_{i=1}^{n} \Gamma^i)(\mathbf{a}), \quad (2.2:92)$$

hence, the optimization with the use of canonical genetic algorithm is often more difficult than the original optimization of the objective function F: $S \rightarrow \Re$. Therefore, more complex representations of chromosomes are proposed, which are more similar to the representation of objects in the original search space.

In the last two decades, there is growing interest in real-valued representation of chromosomes. Many practical applications of parametric optimization uses this representation and indicates its usefulness (Fogel 1997b, Cyran and Mrózek 2001) and greater effectiveness as compared to binary representation (Michalewicz 1992). However, such conclusion, although confirmed by a number of experimental studies, is contradicting the classical interpretation of the fundamental theorems about canonical genetic algorithms. In particular this is the case with the interpretation of the Schema Theorem. This theorem has been formulated for arbitrary finite alphabet, and below, this general version is presented.

Theorem 2.2:4 (The Schema Theorem – after Radcliffe 1997)

Let ξ be the schema over the representation space I. Moreover, let this space be searched by evolutionary algorithm using proportional selection and classical operations of mutation and recombination. Let us also denote by $N_\xi(t)$ the number of schemas ξ in the generation t. Then the number $N_\xi(t+1)$ of this schema in the next generation is given by

$$E\{N_\xi(t+1)\} \geq N_\xi(t)\frac{\hat{f}_\xi(t)}{\bar{f}(t)}\left[1 - D_c(\xi)\right]\left[1 - D_m(\xi)\right] \quad (2.2:93)$$

where $\hat{f}_\xi(t)$ is the observed fitness of the schema ξ in the generation t

$$\hat{f}_\xi(t) = \frac{1}{N_\xi(t)}\sum_{x\in\xi} f(x), \quad (2.2:94)$$

$\bar{f}(t)$ is the average fitness in that generation, whereas $D_c(\xi)$ and $D_m(\xi)$ are the upper limits of the destructive effects on the number of elements belonging to the schema ξ, caused by the recombination and mutation, respectively.

Proof

It is clear that without mutation and recombination the expected number of representatives of schema ξ in generation t+1, $E[N_\xi(t+1)]$, is equal to the number of schema representatives in generation t, $N_\xi(t)$, multiplied by a the relative fitness of that schema $\hat{f}_\xi(t) / \bar{f}(t)$. Mutation and recombination can destroy the schema, however it is very hard to estimate the actual destructive effects of these operations. Rather the upper limits of them are used $D_c(\xi)$ and $D_m(\xi)$ (which are easily computable) and therefore the actual $E[N_\xi(t+1)]$ can be larger than the expression on the right side of formula (93) since the actual destructive effects are typically smaller than their upper limits used in this formula.

■

The Schema Theorem expresses the fact that the number of those schemas which are short (i.e with low recombination destructive effect), low-order (i.e. with low mutation destructive effect) and have the over-average fitness, is exponentially increasing in the population during evolution. Such schemas are referred to as building blocks. Therefore, the hypothesis has been formulated that the evolutionary algorithms are processing not only the chromosomes, but also they implicitly process schemas, which represent chromosomes included in a population.

Assuming the same number of possible solutions represented in the chromosome, coding of these solutions using alphabet {0, 1} assures maximum number of schemas as compared to any other alphabet A, for which *card* (A) > 2. At first it seems that binary alphabet is the most efficient, since it assures the maximum number of schemas to be processed, and therefore the level of hidden parallelism is as big as theoretically possible (Bäck 1997b). However currently, it is often raised that the building blocks hypothesis, being the foundation for the notion of hidden parallelism, is not explaining correctly the mechanism of optimization with the use of genetic algorithms. Additionally, since practical experiments do not confirm better efficiency of binary chromosomes (and even they suggest contrary), therefore, the implications of the Schema Theorem have to be carefully reconsidered. At least these implications should not lead to the conclusion about superiority of the search in binary representation spaces $I = \{0,1\}^l$ (Radcliffe 1997).

Another representations used in evolutionary computation include permutations, finite state machines, trees, neural networks, and others. Permutations are predominantly used in combinatorial problems, which belong to NP- complete problems, and therefore to solve them the heuristics are often used. The example is a traveling salesman problem, which was tried to be solved with the evolutionary approach (Whitley 1997a). To make it effective the

operations of mutation and recombination working in permutation representation space have been proposed.

Whitley (1997b) shows that mutation operator can be implemented as so called 2-opt operator used for local searching. This operator chooses two points along the permutation chain, and then it reverses the sequence in the chosen segment (Fig. 3).

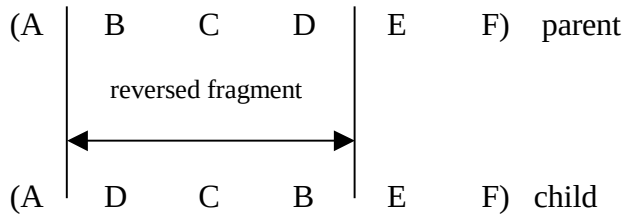


Fig. 2.2:3. Mutation for permutation representation, implemented as 2-opt operator
Rys. 2.2:3. Mutacja dla reprezentacji permutacyjnej implementowana jako operator 2-opt

The recombination can be implemented with the use of operator referred to as crossing-over with ordering (Whitley 1997c). First, the two cutting positions are randomly chosen, and then the genes of the first parent, which are between the cutting points are copied to the child. Finally, starting from the position directly after the second cutting point, the genes of the second parent are examined if they are not present in already created part of the child. If this is satisfied, these genes are copied onto subsequent positions. After reaching the end of the chromosome, the process is continued starting from the first position of the second parent, and continued until the first cutting point. Defined recombination operator inherits from the first parent information about the sequence, absolute position, and adjacencies of genes. However, it inherits only information about sequence form the second parent.

Let us now consider the finite-state representations used in evolutionary computing.

Definition 2.2:22 (Finite state machines, after Fogel 1997d)

The finite state machines are defined as the ordered 5-tuples

$$M = (Q, \tau, \rho, s, o), \tag{2.2:95}$$

where Q is a finite set, called set of states, τ is a finite set of input symbols, ρ is a finite set of output symbols, s is a next state function defined as

$$s : Q \times \tau \to Q, \tag{2.2:96}$$

and o is an output function given by

$$o : Q \times \tau \to \rho, \tag{2.2:97}$$

—

The practical example of finite-state machines is the digital Mealy's machine (Stanczyk et al. 2007), for which Q, τ, ρ are sets of binary vectors (Fig. 4). The mutation operation can be

implemented as a change of the output symbol, a change of the transition between states, adding a new state, deletion of a state, or change of the initial state. The recombination operators are changing particular states of the parent machines. Typical such operators are given for example by Birgmeier (1996).

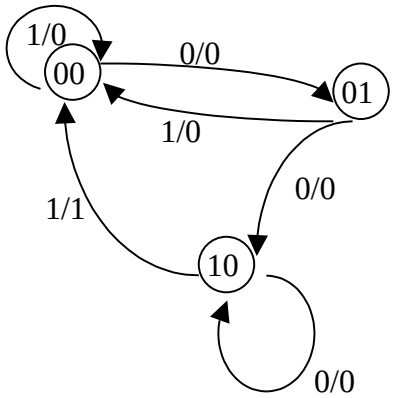


Fig. 2.2:4. The Mealy's machine as the example of finite-state machine
Rys. 2.2:4. Automat Mealy'ego jako przykład maszyny stanów skończonych

In a search for structures representing executable programs or functions, the representation in a form of a parse trees has been proposed (Angeline 1997a). The language defining a program, which is appropriate for the parse trees, should be of the homogeneous type. It means that the values returned by all nodes should of the same type. Not going into details of defining such languages for complex problems, below the possible genetic operators (mutation and recombination) are shown for parse trees representing logic functions.

Mutation can be defined as one of the following operators: switch, cycle, shrink, and grow. Switch operator randomly chooses two nodes of the tree and changes the sub-trees for which the chosen nodes are the roots. (Fig. 5).

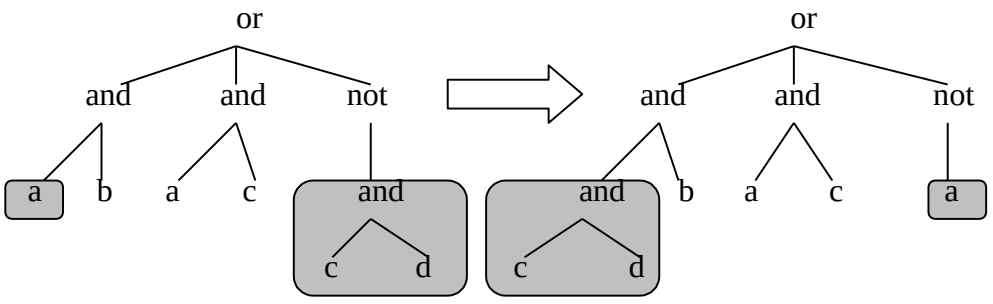


Fig. 2.2:5. Switch operator (after Angeline 1997b)
Rys. 2.2:5. Operator przełączenia (na podstawie Angeline 1997b)

The cycle operator randomly chooses one node in a tree and then randomly changes this node value by other value, representing some operations with the same number of arguments as the original operations (Fig. 6).

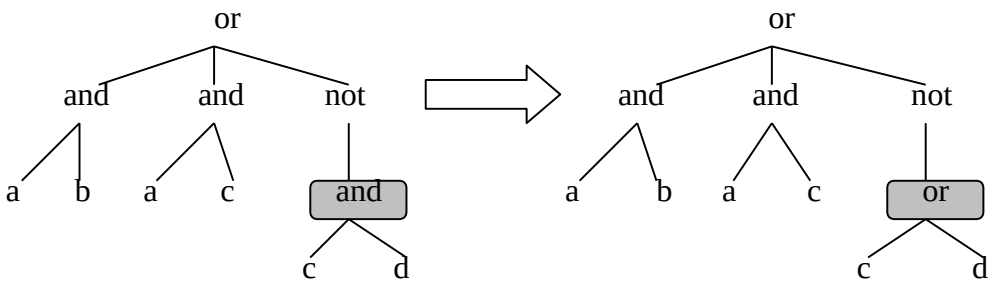


Fig. 2.2:6. Cycle operator (after Angeline 1997b)
Rys. 2.2:6. Operator cykliczny (na podstawie Angeline 1997b)

The shrink operator randomly chooses one non-leaf node of the tree and subsequent change of the sub-tree, of which the chosen node is a root, by one of the leaves of this sub-tree (Fig. 7).

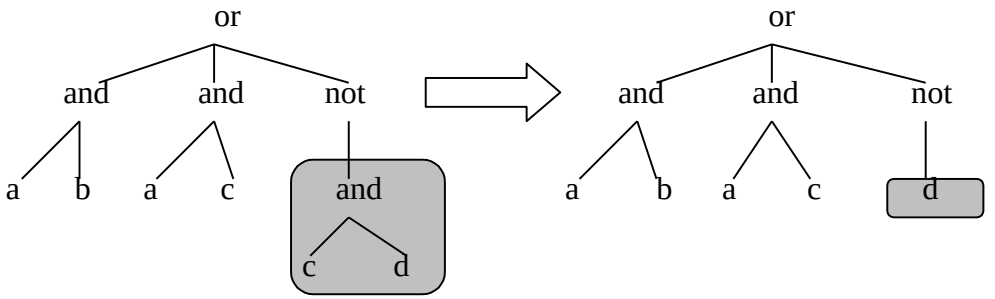


Fig. 2.2:7. Shrink operator (after Angeline 1997b)
Rys. 2.2:7. Operator kurczenia (na podstawie Angeline 1997b)

The grow operator randomly chooses one of the leaves of the tree and then randomly generates sub-tree, in which the chosen leaf is also a terminal element (Fig. 8).

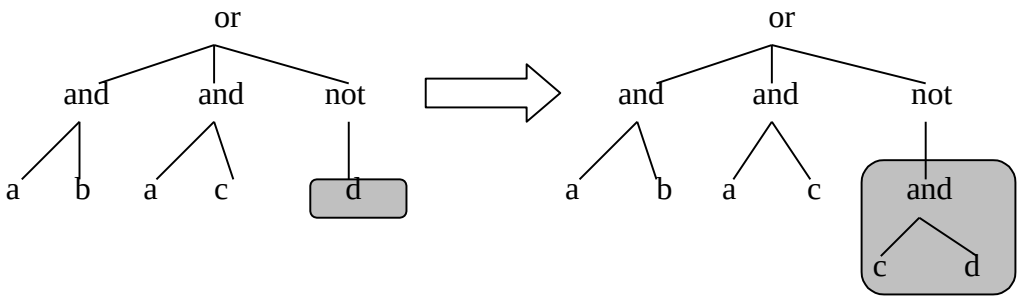


Fig. 2.2:8. Grow operator (after Angeline 1997b)
Rys. 2.2:8. Operator wzrostu (na podstawie Angeline 1997b)

The recombination operator appropriate for this representation randomly chooses two nodes in each of the two parents and subsequent change of the sub-trees between parents (Fig. 9).

Note, that all mentioned above operators create daughter objects, which are syntactically correct representations of the chromosomes, i.e. they are still parse trees. De Jong et al. (1997) discuss also representations which correspond to the sets of rules, programs written in Lisp, or ANNs. This latter representation is considered also by Cyran et al. (1997).

Discussed representations can be used for storing genetic material of haploid individuals. Diploid representations are also possible, which additionally can account for sex of individuals, but, as more complex, they are relatively rarely used, except for modeling actual diploid populations.

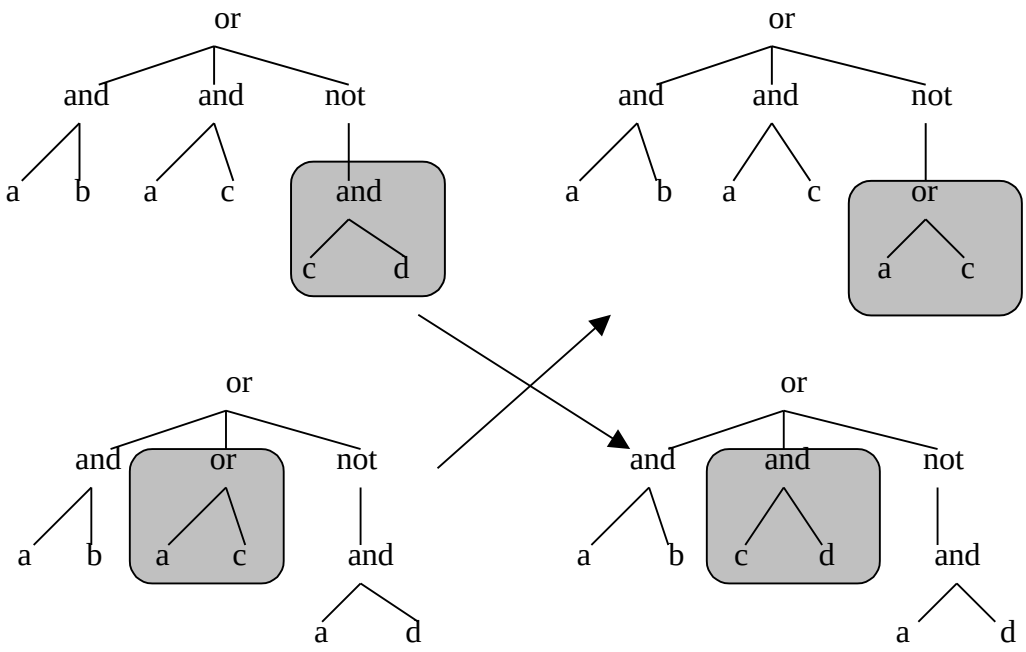


Fig. 2.2:9. Recombination operator for parse trees (after Angeline 1997b)
Rys. 2.2:9. Operator rekombinacji dla drzew wywodu (na podstawie Angeline 1997b)

2.3. Rough sets

The notion of a rough set has been defined for a representation, processing and understanding of imperfect knowledge. Such knowledge must be often sufficient in controlling, machine learning or pattern recognition. The rough approach is based on an assumption that each object is associated with some information, describing it not necessarily in an accurate and certain way. Objects described by the same information are not discernible. The indiscernibility relation, introduced here in an informal way, expresses the

fact that the theory of rough sets does not deal with individual objects, but with classes of objects which are indiscernible. Therefore the knowledge represented by classical rough sets is granular (Pawlak 1991).

The simple consequence of this fact is that objects with natural real-valued representation, hardly match that scheme, and some preprocessing has to be performed, before such objects can be considered in a rough set-based frame. This preprocessing has the goal in making "indiscernible" objects which are close enough (but certainly discernible) in real-valued space. In majority of applications of the rough set theory, this is obtained by subsequent discretization of all real-valued attributes. This, highly nonlinear process, is not natural and disadvantageous in many applications (for example, such as the application presented in section 2.4). However, before an alternative way of addressing the problem will be presented in section 2.3.2, formal definitions of information system and classical indiscernibility relation are given below.

Definition 2.3:1 (Information system, after Pawlak 1982)

The information system S is defined as a 5-tuple $S = <U, Q, v, f>$ composed of a nonempty finite set called universe U, nonempty finite set of attributes Q, function f called the information function, and a mapping v, which associates each attribute $q \in Q$ with its domain V_q.

▬

Definition 2.3:2 (Information function, after Pawlak 1982)

The information function $f : U \times Q \rightarrow V$ is defined in such a way, that $f(x, q)$ reads as the value of attribute q for the element $x \in U$, and V denotes a domain of all attributes $q \in Q$, defined as a union of all domains of single attributes, *i.e.* $V = \bigcup_{q \in Q} V_q$.

▬

Definition 2.3:3 (Classical indiscernibility relation, after Pawlak 1982)

Each nonempty set of attributes $C \subseteq Q$ defines in the information system S the indiscernibility relation $I_0(C) \subseteq U \times U$, given as

$$x \; I_0(C) \; y \quad \Leftrightarrow \quad \forall q \in C, \;\; f(x,q) = f(y,q), \tag{2.3:1}$$

where $x, y \in U$.

▬

Note, that the relation $I_0(C)$ is the equivalence relation, since it is reflexive, symmetric, and transitive (Bolc et al. 1995). Therefore it divides the universe U on abstract classes, what makes rough sets and their extensions appropriate tool for classification (Pawlak and Skowron 2007a, 2007b, 2007c). The family of all abstract classes in this partition is denoted

by C^*, and the particular abstract class of the relation $I(C)$ which contains element $x \in U$ is denoted by $[x]_{I(C)}$.

The definition of indiscernibility relation by equation (1), although theoretically applicable for both, discrete and continuous domains V, is practically valuable only for discrete domains. For continuous domains such relation is too strong, because in practice all elements would have been discernible. Consequently, all abstract classes generated by I_0, would have been composed of exactly one element, what would have made the application of rough set theory notions possible, but senseless.

In concordance with the paradigm assumed by cognitive sciences, any knowledge is associated with the ability of classification of the considered objects or phenomena (elements of the universe). Therefore, it is possible to associate formally (see Pawlak 1995b), the knowledge with information system $S = <U, Q, v, f>$

Definition 2.3:4 (Knowledge, after Pawlak 1995b)

The knowledge K_Q in information system $S = <U, Q, v, f>$ is defined as the partition Q^* generated by the set of attributes in S, what can be written as $K_Q = Q^*$. Moreover, the family of partitions $\{\{q\}^*\}_{q \in Q}$ generated by different attribute subsets C constitutes the knowledge base of the information system $S = <U, Q, v, f>$.

▬

Definition 2.3:5 (Notions, basic notions, and elementary notions, after Pawlak 1995b)

Each subset $X \subseteq U$ is called the notion in S. The basic notions Y in S are notions which are abstract classes of the indiscernibility relation $I_0(\{q\})$ based on single attributes $q \in Q$.

The C-elementary notion Z, called also C-elementary set, is each notion, whose elements $x \in Z$ are C-indiscernible, i.e. they belong to the same abstract class $[x]_{I_0(C)}$ of the relation $I_0(C)$. If $C = Q$, then the C-elementary notion is called the elementary notion in S.

▬

In other words, $Y \in \{q\}^*$, i.e., a basic notion consists of such objects which are indiscernible with respect to single attributes $q \in Q$. On the other hand, for given attribute subset $C \subseteq Q$, such that $card(C) > 1$, it is possible to define C-elementary notions.

Lemma 2.3:1 (after Pawlak 1995b)

The knowledge K_Q generated by $S = <U, Q, v, f>$ is identical to knowledge $K_{Q'}$ generated by $S' = <U, Q', v', f'>$ if and only if their elementary notions are identical.

Proof

Directly from Definition 5 it follows that the elementary notion in information system S, is the abstract class of the relation $I_0(Q)$, which generates the atomic unit of knowledge about

universe U with respect to Q. If notion X is a union, product or complement of elementary notions, then certainly X can be accurately specified using Q, i.e. it is definable using Q. Hence, it is definable with respect to the knowledge K_Q. Obviously, any product, union or complement of definable notions is also definable with respect to knowledge K_Q. In general, knowledge K_Q generates in the information system S all definable notions by deriving them from elementary notions (atomic units of knowledge about universe U). Hence, all what can by accurately expressed using knowledge K_Q, is derived from the elementary notions, and if elementary notions of the information systems $S = <U, Q, v, f>$ generating knowledge K_Q are identical to elementary notions of the information systems $S = <U, Q, v, f>$ generating knowledge $K_{Q'}$ then $S = <U, Q, v, f>$ must be equivalent to $S' = <U, Q', v', f'>$. From equivalency of information systems S and S' it follows that K_Q is identical to $K_{Q'}$.

■

Lemma 2.3:2 (Generality of the knowledge, after Pawlak 1995b)

The knowledge K_Q is more general than the knowledge $K_{Q'}$ if and only if $I_0(Q') \subseteq I_0(Q)$.

Proof

$I_0(Q') \subseteq I_0(Q)$ holds if and only if when each abstract class of the relation $I_0(Q')$ is contained in some abstract class of the relation $I_0(Q)$, but not necessarily the opposite. Therefore, each notion of knowledge K_Q is a combination of some notions of knowledge $K_{Q'}$. Hence, knowledge $K_{Q'}$ is more specific than K_Q, form which it follows that knowledge K_Q is more general than $K_{Q'}$.

▪

As presented above, in the classical theory of rough sets originated by Pawlak (1982, 1991), the indiscernibility relation is generated by the information describing objects belonging to some finite set, called universe. If this information is of discrete nature, than the classical form of this relation is natural and elegant notion. For many applications processing discrete attributes, which describe objects of the universe, such definition of indiscernibility relation is adequate, what implies that area of successful use of classical rough set methodology covers problems having natural discrete representation, consistent with granular nature of knowledge in this theory (Pawlak 1991). Such classical rough set model is particularly useful in automatic machine learning, knowledge acquisition and decision rules generation, applied to problems with discrete data not having enough size for application of statistical methods, which demand reliable estimation of distributions characterizing the underlying process (Mrózek 1992a, 1992b).

If however, the problem is defined in a continuous domain, the classical indiscernibility relation almost surely builds one-element abstract classes, and therefore, it is not suitable for

any knowledge generalization. To overcome this disadvantage, different approaches are proposed. The simplest is the discretization, but if this processes is iterated separately for single attributes, it induces artificial and highly nonlinear transformation of the attribute space.

Other approaches concentrate on the generalization of the notion of indiscernibility relation, postulated to be changed to the tolerance relation (Järvinen 2001, Skowron and Stepaniuk 1996) or similarity relation (Doherty and Szałas 2004, Słowiński and Vanderpooten 1997, 2000). The comparative study focused upon even more general approaches, assuming indiscernibility relation to be any binary reflexive relation, is given by Gomolińska (2002). Another interesting generalization of indiscernibility relation into characteristic relation, applicable for attributes with missing values (lost values or don't care conditions) is proposed by Grzymała-Busse (2003, 2004).

In the section 2.3.2, there is presented the author's modification of classical indiscernibility relation, dedicated for rough set theory applied to real-valued attributes space. Contrary to some other known generalizations described in section 2.3.1, the indiscernibility relation introduced by the author, remains an equivalence relation. This relation is obtained by introducing a structure into a collection of attributes. It defines real-valued subspaces used in a multidimensional cluster analysis, which partition the universe in a more natural way as compared to one-dimensional discretization, iterated in a classical model for each attribute.

Since the classical model is a special case of this modification, the modified version can be considered as more general. But more importantly, it allows for natural processing of real-valued attributes in a rough-set theory, broadening the scope of applications of classical, as well as variable precision rough set model (described in section 2.3.1), since the latter can utilize the proposed modification, equally well. In this way one does not have to resign from the equivalence relation, and, at the same time, one can obtain abstract classes uniting similar objects, which belong to the same clusters in a continuous multidimensional space as it is required by majority of classification problems.

The introduction to the rough set theory in the very classic form, known as the classical rough set approach (CRSA) is followed in section 2.3.1 by the review of some well- known generalizations and modifications. Finally, the original author's generalization of rough sets, applicable to continuous attributes (section 2.3.2), is followed in section 2.3.3 by novel author's modification, called quasi-dominant rough set approach (QDRSA). There is also presented a comparison of its advantages and limitations with other rough set-based methods: classical rough set approach (CRSA) and dominance-based rough set approach (DRSA) using illustrative application discussed further in section 4.2 and 4.3. Such strategy, not only presents theoretical aspects of the types of problems adequate for QDRSA, but also it is able

to demonstrate that the class of problems which can be solved with QDRSA is represented by real world applications, similar to those considered in section 4.3.

2.3.1. Major modifications of rough sets (VPRSM, DRSM, Near sets)

As it has been presented above, since the first publication by Pawlak (1982, 1991) of the rough set theory (RST) as an information retrieval system generating rules, which describes uncertain knowledge in a way alternative to fuzzy sets methodology (Zadeh 1965), many modifications of the RST have been proposed. The most notable of them include Variable Precision Rough Set Model (VPRSM) published by Ziarko (1993), Dominance Rough Set Approach (DRSA) introduced by Greco, Matarazzo and Slowinski (Greco et al. 1999a), and Near Set Theory (NST) developed by Peters (2007).

The first mentioned modification (VPRSM) is dedicated for large data sets, where inconsistencies, tolerated to some extent, can be advantageous. The second (DRSA) is appropriate for attributes with inherent preference order and not necessarily discretized. Finally, the latter (NST), by using affinities between perceptual objects and perceptual granules, provides a basis for perceptual information systems useful in science and engineering. It is also worthwhile to notice that there exists methodology which incorporates Ziarko's idea of variable precision to DRSA methodology resulting in Variable Consistency Dominance Rough Set Approach (VCDRSA) (see Greco et al. 2001).

The crucial notion in the VPRSM is the coefficient describing the level of uncertainty. It specifies, whether the element $x \in U$ belongs to a set $X \subseteq U$ when indiscernible relation $I(C)$ generates the knowledge K_C in information system S.

Definition 2.3:6 (Uncertainty level, after Ziarko 1993)

The uncertainty level coefficient is a function denoted by $\mu_X^C(x)$ and defined as $\mu_X^C(x) = card\ \{ X \cap [x]_{I(C)} \} / card\ \{ [x]_{I(C)} \}$.

▬

Defined above coefficient is also referred to as a rough membership function of an element x, due to similarities with membership function known from the theory of fuzzy sets. This function gave base for the generalization of rough set theory called rough set model with variable precision (Ziarko 1993). This model assumes that lower and upper approximations are dependent on additional coefficient β, such that $0 \le \beta \le 0.5$, and are defined as $\underline{C}_\beta X = \{ x \in U:\ \mu_X^C(x) \ge 1 - \beta \}$ and $\overline{C}_\beta X = \{ x \in U:\ \mu_X^C(x) > \beta \}$ respectively. The boundary in this model is defined as $Bn_C^\beta(X) = \{ x \in U:\ \beta < \mu_X^C(x) < 1 - \beta \}$. It is easy to observe that the classical rough set theory is the special case of variable precision model with $\beta = 0$.

Since $\forall X \subseteq U$, $\underline{C}X \subseteq \underline{C}_{\beta}X \subseteq \bar{C}_{\beta}X \subseteq \bar{C}X$, it follows that VPRSA is a weaker form of the theory as compared to classical model, and therefore, it is often preferable in analysis of large information systems with some amount of contradicting data. The membership function of an element x can be also defined for a family of sets $\mathbf{X}$ as $\mu^{C}_{\mathbf{X}}(x) = card\ \{(\bigcup_{Xn \in \mathbf{X}}, X_n) \cap [x]_{I(C)}\} / card\ \{[x]_{I(C)}\}$. If all subsets X_n of the family $\mathbf{X}$ are mutually disjoint, then $\forall x \in U$, $\mu^{C}_{\mathbf{X}}(x) = \sum_{Xn \in \mathbf{X}}, \mu^{C}_{Xn}(x)$. It is evident that the definition of the rough membership function of the element $\mu^{C}_{X}(x)$ assumes only the existence of classes of equivalence of the relation I, and the VPRSA formally differs from classical model only in the definition of the lower and the upper approximation by the use of this coefficient. Therefore, all rough set-based notions are defined for arbitrary I also in this generalized model.

While the VPRSA was addressing challenges with information processing in large data repositories, the next considered modification of rough set theory, the DRSA, is the response to multicriteria classification problem. The most influential modification encountered in DRSA as compared to CRSA is the change of indiscernibility relation, which is an equivalence relation, to a dominance relation. Using this modification DRSA is able to take into account preference orders in the description of objects by condition and decision attributes.

This is significant improvement, since the well-known methods of knowledge discovery and machine learning do not use the information about preference orders in multicriteria classification. However, taking this information into account can be important in many practical problems, which involve evaluation of objects on preference ordered domains. Therefore, when dealing with such multicriteria classification DRSA often outperforms CRSA, which is not able to make use of this important information – the new model proposed by the author called quasi-dominance rough set approach (refer to Cyran 2009d) addresses the preference order not resigning from the indiscernibility relation, as it is explained in detail in section 2.3.3).

In DRSA like in CRSA, the rough approximation of the partition of information system is a starting point for induction of the IF-THEN decision rules. However, the syntax of these rules is adapted to represent preference orders. The DRSA keeps almost all the best properties of the CRSA: it analyses only facts present in data and possible inconsistencies are not corrected. Moreover, this approach does not need any prior discretization of continuous-valued attributes. In fact, the only known drawback of DRSA is impossibility of using the (relative) value reducts, what motivated the author to propose a hybrid approach, the QDRSA, keeping possibility to use (relative) value reducts and taking into account the

preference order not resigning from the indiscernibility equivalence relation (see section 2.3.3).

Detailed description of DRSA applicable to multicriteria classification and other multicriteria decision problems such as choice and ranking problems is given in Greco et al. (1999b). This latter paper shows that within DRSA heterogeneous information can be effectively processed. The heterogeneity include in this context qualitative and quantitative information, which is ordered and non-ordered and processed using crisp and fuzzy evaluations, as well as ordinal, quantitative and numerical non-quantitative scales of preference.

The applications of DRSA vary from such areas as market analysis, where the usefulness of the DRSA and its advantages over the CRSA are presented on a real study of evaluation of the risk of business failure (Greco et al. 1998) to bioinformatics, where DRSA is applied in the search for signatures of natural selection operating at molecular level (Cyran 2010). Remarkably, DRSA can be applied in conjunction with VPRSM-based concepts, what has been demonstrated by Greco et al. (2001) in Variable Consistency model of DRSA (VCDRSA).

After presenting VPRSM and DRSA, let us focus on NST. This theory, proposed by Peters (2007), was introduced in a context of perception-based approach to studying the nearness of observable objects in a continuum of physical world. The near sets are disjoint sets of such objects that resemble each other, where resemblance between disjoint sets occurs whenever there are observable similarities between the objects in the sets.

In order to determine the similarity between perceptual objects, it is required to compare lists of values, which describe the objects. In other words, a list of such feature values defines an object's description. Hence, comparison of object descriptions provides a basis for NST, whose goal is to offer an efficient framework to group together objects that are perceived as similar based on their descriptions. In particular NST is useful in analysis of digital images perceived as disjoint sets of points (Peters 2009, Peters and Ramanna 2009, Pal and Peters 2010).

The near sets methodology starts with choosing the appropriate method to describe observed objects. This task is accomplished by the selection of probe functions, which represent features of observable objects. Foundations of probe functions were introduced by Pavel (1993). In NST, a notion of a probe function is used as a mapping from an object to a real number, which represents value of an observable feature (Peters 2007). By using probe functions, near sets offer an ideal framework for solving problems based on human perception.

The NST understands perception as a combination of the meaning in psychophysics (Hoogs et al. 2003, Bourbakis 2002) with a view found in Merleau-Ponty's (1945) work.

Psychophysics considers perception of an object, which effects human knowledge about an object, as depending on sense inputs that are the source of signal values, called stimularions, in the cortex of the brain. According to this view, the transmissions of sensory inputs to cortex cells correspond to the probe functions defined in terms of mappings of sets of sensed objects to sets of real-values representing signal values.

This view assumes that the magnitude of each cortex signal value represents a sensation that is a source of object feature values assimilated by the mind. It is based on observation that perception in animals can be modeled as a mapping from sensory cells to brain cells. In particular, visual perception is modeled as a mapping from stimulated retina sensory cells to visual cortex cells. Such mappings, representing probe functions, measure observable physical characteristics of objects in the environment. Therefore a probe function in NST provides a basis for what is commonly known as feature extraction (Guyon et al. 2006) since the sensed physical characteristics of an object can be clearly identified with object characteristic features.

When considering modifications and improvements of the classical rough set approach (CRSA) defined by Pawlak (1991) it may be of some interest to discuss the relation between the given enhanced approach and the original CRSA. Basically there are two kinds of this relation: the first is when the modified approach is more general than the CRSA and then the CRSA is a special case of it, and the second is when the modified approach uses the inspiration from CRSA but in fact it defines a new methodology which cannot be reduced to the CRSA.

The example of the first type is VPRSM, because CRSA is a special case of VPRSM with precision parameter set to one. Also the modified indiscernibility relation, as defined by Cyran (2008b) is more general than the original one, since the latter is a special case of the first. Contrary to these examples, the DRSA is such enhancement which cannot be reduced to classical rough sets: it is inspired by the notions present in RST, but the introduction of dominance relation for preference-ordered attributes (called criteria) instead of equivalence relation present in CRSA is the reason why CRSA cannot be derived from DRSA as its special case.

In this context, the NST is of special type. On one hand, Peters (2007) has proved that near sets are the generalization of rough sets, as each rough set is a near set, and not each near set is a rough set. On the other, the extension of the approximation space (Peters et al. 2007), which is a fundamental notion for RST practically leads to the resignation from the essence of this notion in NST (Peters and Wasilewski 2009). Therefore, although formally NST is a generalization of RST, in practice, it approaches the information granules from a different perspective, which is more focused on search for affinities with the use of tolerance relation, than on defining the approximation space using the equivalence relation.

There have been proposed also other modifications of RST, mainly changing the equivalence relation to the weaker similarity relation (Słowiński and Vanderpooten 2000), or defining the equivalence relation in continuous attribute space without the need of discretization. Introduction of the structure into the set of conditional attributes together with the application of cluster analysis methodology for this purpose has been proposed by Cyran (2008b). This problem is further described in section 2.3.2. The applicability of the latter modification for the problem which was primarily solved with the use of CRSA (see Cyran and Mrózek 2001) has been demonstrated in the case study presented in section 2.4. It is also worth to say that the domain of possible applications of the modified indiscernibility relation extends to all problems with continuous attributes.

2.3.2. Rough sets with real-valued attributes

If a problem is originally defined for real valued attributes, then, before the rough set theory can be used, some clustering and discretization of continuous attributes should be performed. Let this process be denoted as a transformation described by a vector function $\mathbf{\Lambda}: \Re^{card\,(C)} \rightarrow \{1, 2, \ldots, \xi\}^{card\,(C)}$, where ξ is called the discretization factor. The discretization factor simply denotes the number of clusters covering the domain of each individual attribute $q \in C$. Theoretically, this factor could be different for different attributes, but without the loss of generality, we assume its constancy over the set of attributes. Then, the discretization of any individual attribute $q \in C$, can be denoted as a transformation defined by a scalar function $\Lambda: \Re \rightarrow \{1, 2, \ldots, \xi\}$. In this case, one obtains the classical form of indiscernibility relation, defined as (Cyran and Stańczyk 2007a):

$$x \; I_0(\mathbf{\Lambda}[C]) \; y \quad \Leftrightarrow \quad \forall q \in C, \;\; f(x, \Lambda[q]) = f(y, \Lambda[q]), \tag{2.3:2}$$

Below, it will be shown that majority (however, not all) of notions defined in the theory of rough sets *de facto* do not demand the strong version of indiscernibility relation I_0 defined by equation (1) (or by (2), if the discretization is required). From a formal point of view, what is really important, is the fact, that the indiscernibility relation has to be a relation of equivalence, *i.e.* it must be reflexive, symmetric and transitive. From practical point of view, objects indiscernible in a sense of the rough set theory, should be such objects, which are close in a real-valued space.

Hence, the exact form of the indiscernibility relation, as proposed by the classical theory of rough sets, as well as by its generalization VPRSA, is not actually required to create a coherent logical system. Some researchers (Järvinen 2001, Skowron and Stepaniuk 1996, Doherty and Szałas 2004, Słowiński and Vanderpooten 1997, 2000, Gomolińska 2002) go further in this generalizing tendency, resigning from the requirement of equivalence relation.

However, working with such generalizations is often not natural in problems, such as classification, when notion of abstract classes, inherently involved in equivalence relation, is of great importance. Therefore, the author has proposed such modification of the indiscernibility relation, which is particularly useful in many pattern recognition problems, which deal with a space of continuous attributes and which are defined in terms of equivalence relation.

To introduce formally this modification, let us change the notation of indiscernibility relation to be dependent on a family of sets of attributes, instead of being dependent simply on a set of attributes. By the family of sets of attributes, we understand a subset of a power set, based on the set of attributes, such, that all elements of this subset (these elements are subsets of the set of attributes) are mutually disjoint, and their union is equal to the considered set of attributes. This allows to introduce a structure to, originally unstructured, set of attributes, which the relation depends on.

Let $\mathbf{C} = \{ C_1, C_2, \ldots, C_N \}$ denotes the introduced above family of disjoint sets of attributes $C_n \subseteq Q$ such that unstructured set of attributes $C \subseteq Q$ is equal to the union of members of the family **C**, *i.e.* $C = \bigcup_{Cn \in \mathbf{C}} C_n$. Then, let the indiscernibility relation be dependent on **C** instead of being dependent on *C*. Observe that both **C** and *C* contain the same collection of single attributes, however **C** includes additional structure as compared to *C*. If this structure is irrelevant for the problem considered, it can be simply ignored and one can obtain, as a special case, the classical version of indiscernibility relation I_0. However, it is also possible to obtain other versions of this modified relation for which the introduced structure is meaningful.

Let us denote by *I* (without any subscript) an arbitrary relation, having mentioned above properties, reserving subscripts for denoting particular forms of *I*. The exact form of *I*, defined as I_0 in (1) or (2), is not required for processing the rough information, except for some notions, which will be discussed later.

Definition 2.3:7 (Modified indiscernibility relation, after Cyran 2008b)

The modified indiscernibility relation $I_1(C) \in U \times U$ is such form of a relation *I* (in general different from I_0), which is defined as

$$x \; I_1(\mathbf{C}) \; y \quad \Leftrightarrow \quad \forall C_n \in \mathbf{C}, \;\; Clus(x, C_n) = Clus(y, C_n), \tag{2.3:3}$$

where $x, y \in U$, and $Clus(x,C_n)$ denotes the number of a cluster, that the element *x* belongs to.

▬

Theorem 2.3:1 (Generality of modified indiscernibility relation after Cyran 2008b)

The modified indiscernibility relation I_1 is a generalized version of the classical form I_0 of the indiscernibility relation known in CRSA.

Proof (after Cyran 2008b)

Note, that the cluster analysis is required in continuous vector spaces defined by sets of real valued conditional attributes $C_n \in \mathbf{C}$. Note also, that there are two extreme cases of relation I_1, obtained when family **C** is composed of exactly one set of conditional attributes C, and when family **C** is composed of *card* (C) sets, each containing exactly one conditional attribute $q \in C$. The classical form I_0 of the indiscernibility relation can be obtained as the latter extreme special case of the modified indiscernibility relation I_1, because then clustering and discretization is performed separately for each continuous attribute. Hence,

$$I_0(\mathbf{\Lambda}[C]) \equiv I_1(\mathbf{C}) \;\Leftrightarrow\; \mathbf{C} = \left\{ \{q_n\} : C = \bigcup_{q_n \in C} \{q_n\} \right\} \wedge Clus(x, \{q_n\}) = f(x, \Lambda[q_n]). \quad (2.3{:}4)$$

what ends the proof.

■

In other words, the classical form I_0 of the indiscernibility relation can be obtained as a special case of modified version I_1 if we assume that family **C** is composed of such subsets C_n, that each contains just one attribute, and the discretization of each continuous attribute is based on separate cluster analysis as required by a function Λ applied to each of attributes q_n.

One can easily verify (by confrontation of the general form of indiscernibility relation I with presented below notions) that the following constructs form a logically consistent system, no matter what is the specific form of the indiscernibility relation. In particular it is true for such forms of relation I_1 defined by (3), which is different from classical form I_0, defined for discrete and continuous types of attributes in (1) and (2) respectively.

From Definition 9 it follows that set Z is C-elementary, when all elements $x \in Z$ are C-indiscernible, *i.e.* they belong to the same abstract class $[x]_{I(C)}$ of relation $I(C)$. If $C = Q$ then Z is elementary set in S. C-elementary set is therefore the atomic unit of knowledge about universe U with respect to C. Since C-elementary sets are defined by abstract classes of relation I, it follows that any equivalence relation, in particular I_1 can be used as I.

Definition 2.3:8 (C-definable sets, after Pawlak 1991)

If a set X is a union of C-elementary sets then X is C-definable, *i.e.* it is definable with respect to knowledge K_C.

▬

Note, that a complement, a product, or an union of C-definable sets are also C-definable set (notion). Therefore the indiscernibility relation $I(C)$, by generating knowledge K_C, defines all what can be accurately expressed with the use of set of attributes C. Two information systems S and S' are equivalent if they have the same elementary sets. Then the knowledge K_Q is the same as knowledge $K_{Q'}$. Knowledge K_Q is more general than knowledge $K_{Q'}$ iff

$I(Q') \subseteq I(Q)$, *i.e.* when each abstract class of the relation $I(Q')$ is included in some abstract class of $I(Q)$. *C*-definable sets, as unions of *C*-elementary sets are also defined for any equivalence relation *I*.

Definition 2.3:9 (*C*-rough set *X*, after Pawlak 1982, 1991, 1995a)

Any set being the union of *C*-elementary sets is a *C*-crisp set, any other collection of objects in universe *U* is called a *C*-rough set.

▬

A rough set contains a border, composed of elements such, that based on the knowledge generated by indiscernibility relation *I*, it is impossible to distinguish whether or not the element belongs to the set. Each rough set can be defined by two crisp sets, called lower and upper approximation of the rough set. Since *C*-crisp sets are unions of *C*-elementary sets, and *C*-rough set is defined by two *C*-crisp sets, therefore the notion of *C*-rough set is defined for any equivalence relation *I*, in particular for I_1 different than I_0.

Definition 2.3:10 (*C*-lower approximation of rough set $X \subseteq U$, after Pawlak 1982, 1995a)

The lower approximation of a rough set *X* is composed of those elements of universe, which belong *for sure* to *X*, based on indiscernibility relation *I*. Formally, *C*-lower approximation of a set $X \subseteq U$, which is denoted as $\underline{C}X$, is defined in the information system $S = <U, Q, v, f>$ as $\underline{C}X = \{ x \in U: [x]_{I(C)} \subseteq X \}$.

▬

Definition 2.3:11 (*C*-upper approximation of rough set $X \subseteq U$, after Pawlak 1982, 1995a)

The upper approximation of a rough set *X* is composed of those elements of universe, which *perhaps* belong to *X*, based on indiscernibility relation *I*. Formally, *C*-upper approximation of a set $X \subseteq U$, denoted as $\overline{C}X$ is defined in the information system $S = <U, Q, v, f>$ as $\overline{C}X = \{ x \in U: [x]_{I(C)} \cap X \neq \emptyset \}$.

▬

Definition 2.3:12 (*C*-border of rough set $X \subseteq U$, after Pawlak 1982, 1995a)

The border of a rough set is the difference between its upper and lower approximation. Formally, *C*-border of a set *X*, denoted as $Bn_C(X)$ is defined as $Bn_C(X) = \overline{C}X - \underline{C}X$.

▬

Definition 2.3:13 (*C*-positive region of the set $X \subseteq U$, after Pawlak 1991)

C-positive region of a set *X*, *i.e.* such region whose elements can be classified as *for sure* belonging to *X*, is denoted as $Pos_C(X)$ and defined as *C*-lower approximation of *X*.

▬

Definition 2.3:14 (*C*-negative region of the set $X \subseteq U$, after Pawlak 1991**)**

C-negative region of *X*, denoted as $Neg_C(X)$, contains all elements of universe *U*, which *for sure* do not belong to *X*, *i.e.* In other words, it is a complement of *C*-upper approximation of *X*, $Neg_C(X) = U - \bar{C}X$.

▬

Note, that both $\underline{C}X = \{ x \in U: [x]_{I(C)} \subseteq X \}$ and $\bar{C}X = \{ x \in U: [x]_{I(C)} \cap X \neq \emptyset \}$ are *C*-crisp sets, so they can be defined for arbitrary equivalence relation *I*, such as for example relation I_1. Moreover, since $Bn_C(X)$ is a difference of two *C*-crisp sets, its definition is also based on arbitrary equivalence relation *I.* The positive region (similarly like *C*-lower approximation of *X*) can be defined for arbitrary *I*, and the negative region, as a difference of *U* (which does not depend on *I*) and a *C*-crisp set $\bar{C}X$, is also based on arbitrary relation of equivalence *I*.

Note also, that the indiscernibility relation *I* generates in any information system *S* some topology which describes four different topological types of rough sets. These types are: sets roughly definable, sets internally indefinable, sets externally indefinable and sets totally indefinable.

Definition 2.3:15 (Sets roughly *C*-definable, after Pawlak 1995a)

Set *X* is roughly *C*-definable iff $Pos_C(X) \neq \emptyset$ and $Neg_C(X) \neq \emptyset$, *i.e.* universe *U* contains some elements which for sure belong to *X* and some element which for sure do not belong to *X*.

▬

Definition 2.3:16 (Sets internally *C*-indefinable, after Pawlak 1995a)

Rough set *X* is called internally *C*-indefinable iff its positive region is empty, but negative region is not empty, *i.e.* when $Pos_C(X) = \emptyset$ and $Neg_C(X) \neq \emptyset$.

▬

Definition 2.3:17 (Sets externally *C*-indefinable, after Pawlak 1995a)

Rough set *X* is called externally *C*-indefinable iff its positive region is not empty, but negative region is empty, *i.e.* when $Pos_C(X) \neq \emptyset$ and $Neg_C(X) = \emptyset$.

▬

Definition 2.3:18 (Sets totally *C*-indefinable, after Pawlak 1995a)

Rough set *X* is called totally *C*-indefinable iff both positive and negative regions of *X* are empty, *i.e.* when $Pos_C(X) = Neg_C(X) = \emptyset$.

▬

It is easy to observe, that all notions defined in Definitions 15-18 as being declared by specific positive and negative regions, can be defined for any relation *I*, in particular for modified relation I_1.

Notions of a rough set theory, applicable for a separate set *X*, are generally applicable also for families of sets $\mathbf{X} = \{ X_1, X_2, \ldots, X_N \}$, where $X_n \subseteq U$, and $n = 1, \ldots, N$. Examples of such notions are given below in Definitions 19-23.

Definition 2.3:19 (*C*-lower approximation of family of sets, after Mrózek 1998)

The lower approximation of a family of sets is a family of lower approximations of sets belonging to family considered. Formally, $\underline{C}\mathbf{X} = \{\underline{C}X_1, \underline{C}X_2, \ldots, \underline{C}X_N\}$.

▬

Definition 2.3:20 (*C*-upper approximation of family of sets, after Mrózek 1998**)**

The upper approximation of a family of sets is a family of upper approximations of sets belonging to family considered. Formally, $\bar{C}\mathbf{X} = \{\bar{C}X_1, \bar{C}X_2, \ldots, \bar{C}X_N\}$.

▬

Definition 2.3:21 (*C*-border of family of sets, after Mrózek 1998**)**

The boundary of the family of sets **X** is a union of boundaries of sets belonging to the family considered, *i.e.* $Bn_C(\mathbf{X}) = \bigcup_{Xn \in \mathbf{X}}, Bn_C(X_n)$.

▬

Definition 2.3:22 (*C*-positive region of the family of sets, after Mrózek 1998)

The positive region of the family of sets **X** is a union of positive regions of sets belonging to the family considered, *i.e.* $Pos_C(\mathbf{X}) = \bigcup_{Xn \in \mathbf{X}}, Pos_C(X_n)$.

▬

Definition 2.3:23 (*C*-negative region of the family of sets, after Mrózek 1998)

The negative region of the family of sets **X** is defined as $Neg_C(\mathbf{X}) = U - \bigcup_{Xn \in \mathbf{X}}, \bar{C}X_n$.

▬

Note, that concepts defined in Definitions 19-23, as families, differences and unions of *C*-crisp sets, are based on arbitrary relation of equivalence *I*. The theory of rough sets not only defines, as presented above, a framework of coherent notions, used for representation of uncertain knowledge, but also gives tools for associating the objects with numerical uncertainty measures.

Therefore, it follows that the last class of concepts considered in various models of the rough set theory described in this monograph, is a class of coefficients which indicate the accuracy and the quality of the approximation space.

Definition 2.3:24 (*C*-accuracy of approximation of a set: $\alpha_C(X)$, after Pawlak 1995a)

C-accuracy of approximation of a nonempty set *X*, denoted as $\alpha_C(X)$, is given by the ratio of lower and upper approximation of *X*, *i.e.*, $\alpha_C(X) = card\,[Pos_C(X)] / card\,(\bar{C}X)$.

▬

The accuracy of approximation defined in Definition 24 satisfies $0 \leq \alpha_C(X) \leq 1$. Using this coefficient, it is possible to give alternative definitions of crisp and rough sets, as presented in Definition 25. Another coefficient measuring the uncertainty in rough set theory is called quality of approximation defined in Definition 26.

Definition 2.3:25 (Roughness of a set, after Pawlak 1995a)

When $\alpha_C(X) = 1$ then the considered set *X* is *C*-crisp in a system with knowledge K_C generated by the indiscernibility relation $I(C)$. Similarly, if $\alpha_C(X) < 1$, then *X* is called *C*-rough set.

▬

Definition 2.3:26 (*C*-quality of approximation of a set: $\gamma_C(X)$, after Pawlak 1995a)

C-quality of approximation of a set *X*, denoted as $\gamma_C(X)$ is defined as $\gamma_C(X) = card\,[Pos_C(X)] / card\,(U)$.

▬

Interesting comparison of *C*-quality of approximation and Dempster-Shafer theory of evidence is given by Skowron and Grzymała-Busse (1994). In the context considered here, it is important to observe that the notions presented in Defintions 24-26 as numerical ratios of numbers associated with notions defined for any *I*, they are also meaningful for arbitrary relation *I*. Other notions, which are based on a notions of the upper and/or the lower approximation of a family of sets **X**, with respect to a set of attributes *C*, include: *C*-accuracy of approximation of a family of sets, *C*-quality of approximation of a family of sets. This latter coefficient is especially interesting for the application presented in the subsequent section, since it is used as an objective function in a procedure of optimization of the feature extractor. For this purpose, the considered family of sets is a family of abstract classes generated by the decision attribute *d* being the class of the image to be recognized (see. Section 2.4 for the exemplary application). Here, let us define this coefficient for an arbitrary family of sets **X**.

Definition 2.3:27 (*C*-quality of approximation of a family of sets **X**, after Mrózek 1998)

C-quality of approximation of a family of sets **X**, denoted by $\gamma_C(\mathbf{X})$ is defined as $\gamma_C(\mathbf{X}) = card\,[Pos_C(\mathbf{X})] / card\,(U)$.

▬

The analysis of concepts presented above indicates, that they do not require any particular form of the indiscernibility relation (like for example the classical form referred to as I_0). They are defined for any form of the indiscernibility relation (satisfying reflexity, symmetry and transitiveness), denoted by I (in particular I_1, which is very usful in continuous space) and therefore they are strict analogs of classical notions defined with the assumption of original form of indiscernibility relation I_0 defined by equations (1, 2).

Finally, let us discuss some of the notions of rough set theory that cannot be used in a common sense with the modified indiscernibility relation I_1 defined by (3). Let us start with, the so called, basic sets which are abstract classes of relation $I(\{q\})$ defined for singe attribute q. These are simply sets composed of elements indiscernible with respect to single attribute q. Obviously, this notion loses its meaning when I_1 is used instead of I_0, because abstract classes generated by $I_0(\{q\})$ are always unions of some abstract classes generated by $I_0(C)$, however abstract classes generated by $I_1(\{q\})$ not necessarily are unions of abstract classes generated by $I_1(C)$. Therefore the conclusion that knowledge $K_{\{q\}}$ generated by $I_0(\{q\})$ is always more general than knowledge K_C generated by $I_0(C)$, no longer holds when I_1 is used instead of I_0.

Similarly, notions of reducts, relative reducts, cores and relative cores no longer are applicable in their classical sense, since their definitions are strongly associated with single attributes. Joining these attributes into members of family **C**, destroys the individual treatment of attributes, required for these notions to have their well known meaning. However, as long as the rough set theory is used in the continuous attribute space, to the extent not going beyond notions described Definitions 8-27, the modified I_1 version should be considered more advantageous, as compared to the classical form I_0. In particular, this is true in processing of the knowledge obtained from the holographic ring-wedge detector (given as the illustrative example in section 2.4), when the quality of approximation of family of sets plays the major role in the quality of recognition.

2.3.3. Quasi dominance rough set approach

In this section the novel methodology developed by the author (Cyran 2009d), called quasi dominance rough set approach (QDRSA) is presented. QDRSA can be considered as a hybrid of classical rough set approach (CRSA) and dominance rough set approach (DRSA). After presenting this methodology, the advantages of QDRSA over CRSA and DRSA are illustrated for certain class of problems together with limitations of proposed methodology for other types of problems where CRSA or DRSA are better choice. The analysis of the reasons why QDRSA can produce decision algorithms yielding smaller error rates than DRSA is performed on the real world example, presented in section 4.3.3. This example shows that superiority of QDRSA over CRSA and DRSA in certain types of applications is of practical value.

The DRSA is claimed to have many advantages over CRSA in applications with natural preference-ordered attributes. Not denying this statement in general, it is possible to demonstrate the example of such information system S with preference-ordered attributes, which, when treated as a decision table, can yield better (in the sense of decision error) decision algorithm A than that generated by DRSA (A_{DRSA}).

The superiority of algorithm A is also true (however in the sense of larger generality level) when the aforementioned algorithm A is compared with the algorithm A_{CRSA} obtained by application of CRSA. The quasi dominance rough set approach is the framework within which the algorithm A can be derived. That is why the algorithm A will be referred to as A_{QDRSA}.

The QDRSA can be considered as a hybrid of CRSA and DRSA. Like DRSA it is dedicated for problems with preference-ordered attributes, but contrary to DRSA, it does not resign from the classical indiscernibility.

Definition 2.3:28 (Indiscernibility relation in QDRSA, after Cyran 2009d)

For the information system $S = (U, \underline{Q}, V_q, f)$ in which $Q = C \cup \{d\}$ and for any $x, y \in U$ the I_{QDRSA} is defined as

$$x \; I_{QDRSA}(\mathrm{C}) \; y \quad \Leftrightarrow \quad \forall q \in C, \;\; f(x,q) = f(y,q). \tag{2.3:5}$$

■

Comparison of formula (1) in Definition 3 with formula (5) in Definition 28 reveals that the eqivalence relations I_0 and I_{QDRSA} are identical. Therefore, the notions of lower and upper approximations, as well as particularly important for classification notions of quality of approximation, (relative) cores, (relative) reducts and (relative) value reducts are defined in QDRSA like in CRSA.

In particular, it follows in QDRSA that an attribute $q \in C$ is redundant in C, where $C \subseteq Q$, if the indiscernibility relation $I_{QDRSA}(B)$ is identical to the indiscernibility relation $I_{QDRSA}(C - \{q\})$ or, what is equivalent, if the attribute q is functionally dependent of the subset $C - \{q\}$, what can be denoted as $C - \{q\} \rightarrow q$. If $I_{QDRSA}(C) \neq I_{QDRSA}(C - \{q\})$ then the attribute $q \in C$ is irredundant in C (i.e. it is irremovable from C). Set of attributes $C \subseteq Q$ is independent if each attribute $q \in C$ is irredundant in C. Otherwise, a set of attributes $C \subseteq Q$ is dependent.

Definition 2.3:29 (Reduct, after Pawlak 1995a, adapted to QDRSA)

In QDRSA, similarly like in CRSA, each set $Q' \subseteq Q$ is a reduct of the set Q in the information system $S = <U, Q, v, f>$ if Q' is independent and if $I_0(Q') = I_0(Q)$.

■

It follows that reduct *Q'* is the smallest (in the sense of sets inclusion) subset of attributes which generates the same classification of the elements in the universe *U*, as the complete set of attributes *Q* does. At the same time, the reduct *Q*' is the largest (in the sense of sets inclusion) independent subset of the set *Q*. So, the attributes not belonging to the reduct *Q*', as being dependent of attributes of this reduct, are redundant for the classification of elements in the universe *U*. In given information system there could be many reducts, and moreover, it is possible to define not only a reduct of the complete set of attributes *Q*, but also reducts of some subsets $C \subseteq Q$.

Denote the set of all reducts of the set *C* in the information system *S* by *RED* (*C*). Then, it follows that (Mrózek 1998)

$$\forall P \subseteq Q, C \subseteq Q, C' \in RED(C): \left(C \xrightarrow{k} P\right) \Rightarrow \left(C' \xrightarrow{k} P\right), \tag{2.3:6}$$

where $C \xrightarrow{k} P$ denotes that a set of attributes $P \subseteq Q$ depends at the k^{th} level ($0 \le k \le 1$) on the set of attributes $C \subseteq Q$. This latter means that for k^{th} fraction of all elements of the universe *U* the values of attributes from *P* can be reconstructed having values of attributes from *C*. Moreover, the following statements are also true (Mrózek 1998)

$$\forall C \subseteq Q, C' \in RED(C): \; C' \to C - C', \tag{2.3:7}$$

$$\forall C \subseteq Q, C' \in RED(C), p \in C', q \in C': \; \neg(\{p\} \to \{q\}) \wedge \neg(\{q\} \to \{p\}), \tag{2.3:8}$$

Theorem 2.3:2 (Reduct-based knowledge)

Assuming that *Q*' is a reduct of *Q*, the knowledge $K_Q = Q^*$ contained in the information system $S = <U, Q, v, f>$ considered in QDRSA is identical to the knowledge $K_{Q'} = (Q')^*$ contained in the information system $S' = <U, Q', v, f>$ derived from *S* by reducing *Q* to *Q'*.

Proof

Any reduct *Q'* of the set of attributes *Q* is such minimum subset of *Q*, which generates identical set of elementary notions in the information system *S*. Therefore, knowledge $K_Q = Q^*$ contained in the information system $S = <U, Q, v, f>$ is based on the same set of elementary notions as knowledge $K_{Q'} = (Q')^*$ contained in the information system $S' = <U, Q', v, f>$. Hence, based on Lemma 1 valid for CRSA, and using identity of I_0 in CRSA with I_{QDRSA} in QDRSA, it follows that knowledge K_Q is identical to knowledge $K_{Q'}$.

■

QDRSA, similarly to CRSA, also uses notion of the core of attributes, which is related to reducts, as described below.

Definition 2.3:30 (Core, after Pawlak 1995a, adapted to QDRSA)

The core of the set of attributes $C \subseteq Q$ in the information system, denoted by *CORE*(*C*), is a set of all attributes irremovable from *C*

$$CORE(C) = \{q \in C : I(C - \{q\}) \neq I(C)\}, \tag{2.3:9}$$

■

Hence, the core $CORE(Q)$ of the set of attributes Q in the information system $S = <U, Q, v, f>$ defines the knowledge $K_{CORE(Q)} = (CORE(Q))^*$, which cannot be removed in any reduction process, minimizing the size of the original knowledge $K_Q = Q^*$, without loss of classification abilities. Therefore, the knowledge $K_{CORE(Q)}$ is in a sense the most relevant part of the knowledge K_Q, and the core itself is the most relevant subset of attributes. Nevertheless, it is possible that the core $CORE(Q)$ is an empty set and then there is lack of such essential part of knowledge in the information system $S = <U, Q, v, f>$.

The following relation is true between the notion of core and the reduct of the set of attributes (see Pawlak 1995a)

$$\forall C \subseteq Q: \quad CORE(C) = \bigcap_{C' \in RED(C)} C', \tag{2.3:10}$$

Theorem 2.3:3 (Core and reduct relationship)

The core is included in each reduct.

Proof

Based on (10), the core is the intersection of all reducts. Hence

$$\forall q \in Q, C' \in RED(C): \quad q \in CORE(C) \Rightarrow q \in C', \tag{2.3:11}$$

what ends the proof.

■

Like CRSA, the QDRSA is able to exploit the relative counterparts of many concepts. The description of relative independence, relative reducts and relative cores will explain this issue in more detail.

Definition 2.3:31 (Relative independence, after Pawlak 1995a, adapted to QDRSA)

In the information system $S = <U, Q, v, f>$ the attribute set $C \subseteq Q$ is relatively independent with respect to the set of attributes $R \subseteq Q$ (i.e. it is R-independent) if for each proper subset $P \subset C$ the following inequality is satisfied: $Pos_P(R^*) \neq \mathrm{Pos}_C(R^*)$, where $Pos_P(R^*)$ denotes positive region of the family R^* with respect to set of attributes P (see Pawlak 1982, 1991). Otherwise, the set of attributes $C \subseteq Q$ is dependent with respect to set of attributes $R \subseteq Q$ (i.e., it is R-dependent).

■

Note, that for relatively independent set of attributes C, each removal of the attribute from this set results in worse quality of classification of the abstract classes generated by relation $I(R)$ using attributes from C.

Lemma 2.3:3 (Independence)

Classical independence of the set of attributes $R \subseteq Q$ is equivalent to the relative C-independence of the set of attributes $R \subseteq Q$ if $R = C$.

Proof

When $R = C$, the condition of relative independence can be transformed to the condition of classical independence ($I(P) \neq I(C)$):

$$R = C \quad \Rightarrow \quad Pos_P(R^*) \neq Pos_C(R^*) \Leftrightarrow Pos_P(C^*) \neq Pos_C(C^*), \tag{2.3:12}$$

Then, the generalized notion of relative independence becomes classical notion of independence, what ends the proof.

■

Definition 2.3:32 (Relative reduct, after Pawlak 1995a, adapted to QDRSA)

Set $C' \subseteq C$ is called the relative reduct of C with respect to R (R-reduct of C) if C' is R-independent subset of C and $Pos_C(R^*) = Pos_{C'}(R^*)$, or, what is equivalent, if C' is the biggest (in the sense of set inclusion) R-independent subset of C.

▬

Theorem 2.3:4 (Generality of the relative reduct)

The relative reduct is generalized version of the classical reduct.

Proof

If $R = C$ then R-reduct of the set C becomes C-reduct of the set C. From Lemma 3 it follows that independence of the set C is equivalent to the relative C-independence of the same set C. Therefore, C-reduct of the set C becomes the reduct of the set C. Hence, classical reduct is the special case of the relative reduct, what ends the proof.

■

The set of attributes can have more than one relative reduct. Consider the family of all R-reducts of the set $C \subseteq Q$, denoted by $RED_R(C)$. Then it follows that

$$\forall B \subseteq Q, R \subseteq Q, C' \in RED_R(C): \quad Pos_{C'}(R^*) = Pos_C(R^*), \tag{2.3:13}$$

and

$$\forall C \subseteq Q, R \subseteq Q, C' \in RED_R(C): \quad C \xrightarrow{k} R \Rightarrow C' \xrightarrow{k} R. \tag{2.3:14}$$

Definition 2.3:33 (Relative irremovability, after Pawlak 1995a, adapted to QDRSA)

In the in the information system $S = <U, Q, v, f>$ with $C \subseteq Q$ and $R \subseteq Q$, the attribute $q \in C$ is relatively redundant in C (relatively removable form C) with respect to R (R-redundant or R-removable) when $Pos_C(R^*) = Pos_{C-\{q\}}(R^*)$. The attribute $q \in C$ is

relatively irremovable from C with respect to R (R-irremovable), when $Pos_C(R^*) \supset Pos_{C-\{q\}}(R^*)$.

Lemma 2.3:4 (Irremovability)

Relative C-irremovability of the attribute q from the set C is equivalent to classical irremovability of q from C.

Proof

When $R = C$, then the condition of R-irremovability of attribute q from set B becomes the condition $I(C) \neq I(C-\{q\})$ of the classical irremovability of q from C:

$$R = C \quad \Rightarrow \quad Pos_{C-\{q\}}(R^*) \subset Pos_C(R^*) \Leftrightarrow Pos_{C-\{q\}}(C^*) \subset Pos_C(C^*). \qquad (2.3:15)$$

and the notion of R-irremovability of the attribute q from the set C becomes the classical notion of irremovability, what ends the proof.

■

Definition 2.3:34 (Relative core, after Pawlak 1995a, adapted to QDRSA)

The relative core $CORE_R(C)$ of the set of attributes C with respect to R (R-core of the set C) is defined as the set of all R-irremovable attributes from the set of attributes C

$$CORE_R(C) = \left\{q \in C : Pos_C(R^*) \neq Pos_{C-\{q\}}(R^*)\right\}. \qquad (2.3:16)$$

Theorem 2.3:5 (Generality of the relative core)

The relative core is a generalization of the classical core.

Proof

If $R = C$ then R-core of the set C becomes C-core of the set C. Based on Lemma 4, it follows that the classical irremovability of attribute q from the set C is equivalent to relative C-irremovability of the same attribute from the set C. Hence, relative C-core of the set C becomes the core of the set C, what proves that classical core is the special case of the relative core.

■

Summarizing, like in CRSA, in QDRSA, relative reduct and relative core are generalizations of the reduct and the core, respectively, and they relay on relative dependence and independence of attributes. Furthermore, it follows that R-core and R-reduct are satisfying the formula (Pawlak 1985a):

$$CORE_R(C) = \bigcap_{C' \in RED_R(C)} C' . \qquad (2.3:17)$$

It is worth to notice that since the intersection of *R*-reducts can be an empty set, therefore, there exist possibility that the set of attributes does not have the relative core.

Presented above generalizations of the classical notions of core and the reduct are relevant in classification problems, when the information system $S = <U, Q, v, f>$ becomes the decision table $T = <U, C, D, v, f>$ (see Mrózek 1992a) by letting $Q = C \cup D$, i.e. by separating the conditional and decision attributes (C and D respectively). In fact, some special cases of these notions are really important, however these are special cases different that those, which reduced to the classical notions. This problem is explained below in detail.

In the analysis of the decision tables (and therefore in the classification problems such as those considered in the section 2.4 and section 4.3.3) there is used the following special case of the notions defined for the information system $S = <U, Q, v, f>$. Consider two sets of attributes $C, D \subseteq Q$ such that $C \cap D = \varnothing$ and $C \cup D = Q$. Then $S = <U, Q, v, f>$ becomes the decision table $T = <U, C, D, v, f>$ and all conclusions concerning the reduction of the size of knowledge covered in information system can be used for minimization of the number of conditional attributes C in classification problems.

More precisely, for the given decision table $T = <U, C, D, v, f>$, the D-core of the set of conditional attributes C, denoted as $CORE_D(C)$, constitutes the most essential set of attributes from the classification point of view. It includes all these attributes which cannot be removed without reducing the determinism level $\gamma_C(D^*)$ of the decision table T. On the other hand, the D-reduct of the set of conditional attributes C defines in the decision table $T = <U, C, D, v, f>$ such set of the conditional attributes $C' \in RED_D(C)$, which generates the new decision table $T' = <U, C', D, v, f>$ derived from the original table T by cutting C to C' and such that T' is equivalent with T in terms of decision rules covered.

While the notions presented above are defined in the CRSA and QDRSA, after a modification of equivalence relation to tolerance relation (and therefore, after changing abstract classes to dominance cones) these notions are incorporated to the DRSA without the loss of the general meaning. However, there are also notions defined in CRSA and QDRSA which cannot be used in DRSA in their common sense. In fact, existence of such concepts which cannot be directly incorporated to DRSA inspired the author to propose the QDRSA. Within this latter model, there is used the information about the preference order in attribute values (like in DRSA) but (contrary to DRSA) this information is incorporated in such a way, which preserves the equivalence relation, and therefore such concepts as relative value reducts, defined below, can be efficiently utilized.

It follows that in QDRSA (but not in DRSA) further simplification of the information system $S = <U, Q, v, f>$ can be implemented by such elimination of the value of particular attribute for some elements of the universe (however without eliminating the attribute from S)

that the classification ability is not reduced. The notions used in this type of the knowledge reduction are analogues to the notions used in the reduction of redundant attributes.

Definition 2.3:35 (Irremovability for given element, after Pawlak 1995a, adapted to QDRSA)

In the information system $S = <U, Q, v, f>$ with $C \subseteq Q$, the value of the attribute $q \in C$ is removable for the element $x \in U$ if and only if $[x]_{I(C)} = [x]_{I(C-\{q\})}$. Otherwise the value of the attribute q is irremovable for x.

▬

Definition 2.3:36 (Independence for given element, after Pawlak 1995a, adapted to QDRSA)

The set of attributes C is independent for the element x if and only if for each attribute $q \in C$, the value of q is irremovable for x.

▬

Definition 2.3:37 (Value reduct, after Pawlak 1995a, adapted to QDRSA)

The subset $C' \subseteq C$ is a value reduct of the set C for the element $x \in U$ if and only if C' is independent for x and $[x]_{I(C)} = [x]_{I(C')}$.

▬

Definition 2.3:38 (Value core, after Pawlak 1995a, adapted to QDRSA)

For the given element x, the set of all irremovable values of attribute $q \in C$ is referred to as the value core $CORE^x(C)$ of the set C for x.

▬

Note, that there can exist more than one value reduct of the set C for element x. The set of all such value reducts is denoted by $RED^x(C)$. It follows that (Palak 1995a):

$$CORE^x(C) = \bigcap_{C' \in RED^x(C)} C' . \qquad (2.3:18)$$

Analogously to notions of core and reduct, which could be generalized to their relative counterparts, it is possible to generalize notions of value core and value reduct (see below). The special cases of these generalizations are relevant for the reduction of the size of decision algorithm without loss of classification ability of it.

Definition 2.3:39 (Relative irremovability for given element, after Pawlak 1995a, adapted to QDRSA)

For information system $S = <U, Q, v, f>$ and $C \subseteq Q$, the value of attribute $q \in C$ is R-removable for element $x \in U$ if and only if the relation $[x]_{I(C)} \subseteq [x]_{I(R)}$ induces also the relation $[x]_{I(C-\{q\})} \subseteq [x]_{I(R)}$. Otherwise, the value of the attribute q is R-irremovable for x.

▬

Definition 2.3:40 (Relative independence for given element, after Pawlak 1995a, adapted to QDRSA)

Moreover, a set of attributes C is said to be R-independent for the element x if and only if for each attribute $q \in C$ value of q is R-irremovable for x.

▬

Definition 2.3:41 (Relative value reduct, after Pawlak 1995a, adapted to QDRSA)

The subset $C' \subseteq C$ is called the value R-reduct of the set C for the element $x \in U$ if and only if C' is R-independent for x and the relation $[x]_{I(C)} \subseteq [x]_{I(R)}$ induces also the validity of relation $[x]_{I(C')} \subseteq [x]_{I(R)}$.

▬

Definition 2.3:42 (Relative value core, after Pawlak 1995a, adapted to QDRSA)

The set of all values of attribute $q \in C$ which are R-irremovable for the element x is called the value R-core of the set C for x and denoted as $CORE_R{}^x(B)$.

▬

As before, there can be more than one value R-reduct of the set C for the element x. The set of all such value R-reducts us denoted by $RED_R{}^x(C)$. Moreover, it follows that (Pawlak 1995a)

$$CORE_R^x(C) = \bigcap_{C' \in RED_R^x(C)} C' . \tag{2.3:19}$$

Note, that the value R-reduct and the value R-core of the set of attributes C become the classical value reduct and classical value core, respectively, when $R = C$. However, for the classification the more relevant is different special case, namely, when $R \cap C = \varnothing$ and $R \cup C = Q$. These two conditions are satisfied by sets of conditional and decision attributes C and D of the decision table $T = <U, C, D, v, f>$ derived from the information system $S = <U, Q, v, f>$.

Therefore, in a sense, for the given decision table $T = <U, C, D, v, f>$, the value D-core of the set C, denoted as $CORE_D{}^x(C)$, constitutes the set of attributes which are the most relevant in decision making process for the abstract class $[x]_{I(D)}$. It contains all these values of attributes which cannot be removed keeping the determinism level $\gamma_C(D^*)$ of the decision table T unchanged. At the same time, the value D-reduct of the set C defines minimum set of conditional attributes $C' \in RED_D{}^x(C)$, which generates the rule with decisions belonging to the abstract class $[x]_{I(D)}$, such that this rule is equivalent to the decision rule generated by the complete set C for $[x]_{I(D)}$.

Since the QDRSA contrary to DRSA uses the equivalence relation (5), the discernibility matrices can be used for obtaining decision rules. Defined above notions of reduct and core (in different variants) serve for determination of the minimum number of the minimum in size decision rules which are equivalent with respect to the information content to the decisions rules of the original decision table *T*. In order to efficiently compute reducts and the core, the notions of discernibility matrix and discernibility function have been introduced.

Definition 2.3:43 (Discernibility matrix, after Skowron and Rauszer 1992)

In the information system $S = <U, Q, v, f>$, let $C \subseteq Q$, $n = card\,(U)$, and $x_i, x_j \in U$ for $i, j = 1, 2, ..., n$. Then, the discernibility matrix of the set of attributes C, denoted as $\mathbf{M}\,(C) = (m_{ij})$ is symmetrical square matrix of the size $n \times n$, whose elements m_{ij} satisfy

$$m_{ij} = \{q \in C : f(x_i, q) \neq f(x_j, q)\}. \tag{2.3:20}$$

■

The elements m_{ij} of the discernibility matrix are sets of attributes which have different values for objects x_i and x_j. The diagonal of this matrix is composed of the empty sets. Note, that the discernibility matrix $\mathbf{M}\,(C)$ associates with each pair of objects $x, y \in U$ subset of attributes $\delta\,(x, y) \subseteq C$ which satisfies the following properties (see Skowron and Rauszer 1992):

$$\forall x \in U: \quad \delta(x, x) = \varnothing, \tag{2.3:21}$$

$$\forall x, y \in U: \quad \delta(x, y) = \delta(y, x), \tag{2.3:22}$$

$$\forall x, y, z \in U: \quad \delta(x, z) \subseteq \delta(x, y) \cup \delta(y, z), \tag{2.3:23}$$

The above three properties are the properties of the distance in the metric space defined for operators used in the algebra of sets. The function δ is the measure of the distance, and $\delta(x, y)$ is the distance between element x and y in this space. Therefore, the discernibility matrix $\mathbf{M}\,(C)$ can be considered as the distance matrix, because its elements $m_{ij} = \delta\,(x_i, x_j)$ denote the distance between x_i and x_j. It is even better seen if the distance measure in this space is defined as the new function $\delta' = card\,(\delta)$. Then the following properties are satisfied for arithmetical operators (see Skowron and Rauszer 1992):

$$\forall x \in U: \quad \delta'(x, x) = 0, \tag{2.3:24}$$

$$\forall x, y \in U: \quad \delta'(x, y) = \delta'(y, x), \tag{2.3:25}$$

$$\forall x, y, z \in U: \quad \delta'(x, z) \leq \delta'(x, y) + \delta'(y, z), \tag{2.3:26}$$

Compare the Definition 43 defining the discernibility matrix $\mathbf{M}\,(C)$ by equation (20) with Definition 30 defining the core $CORE\,(C)$ by equation (9). Such comparison leads to conclusion that the core can be obtained from the discernibility matrix as set of all those

matrix elements which contain the single attribute (Pawlak 1995a). Formally, this can be denoted as

$$CORE(B) = \{q \in C : (\exists i, j : m_{ij} = \{q\})\}, \tag{2.3:27}$$

On the other hand, the reduct C' of the set of attributes C is the minimum (with respect to the set inclusion) subset of C, such that $C' \cap m_{ij} \neq \varnothing$ for each nonempty element m_{ij} of the matrix $\mathbf{M}(C)$. It is so, because the reduct C' of the set C is such minimum subset of attributes, based on which it is possible to discern all those elements of the universe which are discernible by the whole set C.

In order to compute the relative core $CORE_D(C)$ and the set of the relative reducts $RED_D(C)$ for the decision table $T = <U, C, D, v, f>$ the modified version of the discernibility matrix is required.

Definition 2.3:44 (Modified discernibility matrix, after Skowron and Rauszer 1992)

In the information system $S = <U, Q, v, f>$, with $C \subseteq Q$, $n = card(U)$, and $x_i, x_j \in U$ for $i, j = 1, 2, ..., n$, the modified discernibility matrix, denoted as $\mathbf{M}_D(C) = (m_{ij})$ is symmetrical square matrix of the size $n \times n$, whose elements m_{ij} satisfy

$$m_{ij} = \{c \in C : f(x_i, c) \neq f(x_j, c) \wedge w(x_i, x_j)\}, \tag{2.3:28}$$

where

$$\begin{aligned} w(x_i, x_j) \equiv\ & x_i \in Pos_C(D^*) \wedge x_j \notin Pos_C(D^*) \vee \\ & x_i \notin Pos_C(D^*) \wedge x_j \in Pos_C(D^*) \vee \\ & x_i, x_j \in Pos_C(D^*) \wedge (x_i, x_j) \notin I(D). \end{aligned} \tag{2.3:29}$$

■

If the decision table $T = <U, C, D, v, f>$ is well defined, i.e., when $Pos_C(D^*) = U$, then the condition $w(x_i,x_j)$ in the above definition can be simplified to $(x_i,x_j) \notin I(D)$. It is clear that the element m_{ij} of the matrix $\mathbf{M}_D(C)$ is the set of all conditional attributes $c \in C$, which discern elements x_i and x_j not belonging to the same abstract class of the relation $I(D)$.

The relative core $CORE_D(C)$ can be obtained from the modified discernibility matrix $\mathbf{M}_D(C)$ as a set of all elements of this matrix, which comprise the single attribute (Pawlak 1995a). Formally, it is denoted as

$$CORE_D(C) = \{c \in C : (\exists i, j : m_{ij} = \{c\})\}, \tag{2.3:30}$$

At the same time, the D-reduct C' of the set of attributes C is the minimum (with respect to set inclusion) subset of C, such that $C' \cap m_{ij} \neq \varnothing$ for each nonempty element m_{ij} of the matrix $\mathbf{M}_D(C)$. In other words, the D-reduct C' of the set C is such minimum subset of attributes, based on which it is possible to discern all abstract classes of the relation $I(D)$ which are discernible by the whole set C.

The value reducts and the value core are obtainable from the discernibility matrix **M** (*C*). The value core $CORE^{x_k}(C)$ of the set of attributes $C \subseteq Q$ for the element $x_k \in U$ is the set of all elements m_{ij} of the discernibility matrix **M** (*C*), for which $i = k$ and which comprise single attribute. Formally, it is denoted as

$$CORE^{x_k} C = \{q \in C : (\exists j : m_{kj} = \{q\})\}, \tag{2.3:31}$$

Consequently, the value reduct *C*' of the set of attributes *C* for the element $x_k \in U$ is minimum (with respect to set inclusion) subset of *C*, such that $B' \cap m_{kj} \neq \varnothing$ for each nonempty element m_{kj} in the matrix **M** (*C*).

In other words, the value reduct *C*' of the set *C* for element x_k is the minimum subset of attributes, based on which it is possible to discern x_k from all elements of the universe discernible from x_k by the whole set *C*.

Finally, the relative value core $CORE_D^x(C)$ and the set of relative value reducts $RED_D^x(C)$ can be obtained in a way identical to that, used for classical value core and classical value reducts, respectively, when the modified discernibility matrix $\mathbf{M}_D(C)$ is considered instead of discernibility matrix **M** (*C*).

In practice, the reducts can be determined in the information system $S = <U, Q, v, f>$ using the discernibility function.

Definition 2.3:45 (Discernibility function, after Skowron and Rauszer 1992)

Each discernibility matrix **M** (*C*) uniquely defines the Boolean function called the discernibility function *F* (*C*) given as

$$F(C) = \prod_{(x,y) \in U^2 : \delta(x,y) \neq \varnothing} \left(\sum_{q \in \delta(x,y)} L(q) \right), \tag{2.3:32}$$

where, for any attribute $q \subseteq \delta(x, y)$ the mapping *L* (*q*) denotes uncomplemented Boolean variable $\mathsf{q} \in \{0, 1\}$ uniquely associated with the attribute *q*, whereas Π and Σ denote logical product and sum, respectively.

■

Since all Boolean variables in *F*(*C*) are uncomplemented, the discernibility function is a positive Boolean function. The normal sum form of the discernibility function *F*(*C*) indicates set of all reducts of the set of attributes *C*. Each product term of the function corresponds to a single reduct. This reduct is composed of the attributes $q = L^{-1}(\mathsf{q})$, corresponding to Boolean variables in the given product term. Due to this relationship between normal sum form of the discernibility function *F*(*C*) and the set of all reducts of the set of attributes *C*, the search for reducts is reduced to the search for normal sum form of the positive Boolean function.

The relative reducts for the decision table $T = <U, C, D, v, f>$ are obtainable form the modified discernibility matrix $\mathbf{M}_D(C)$.

Definition 2.3:46 (Relative discernibility function, after Skowron and Rauszer 1992)

Each matrix $\mathbf{M}_D(C)$ uniquely defines the Boolean function called relative discernibility function $F_D(C)$, given by

$$F_D(C) = \prod_{(x,y)\in U^2:\delta(x,y)\neq\varnothing} \left(\sum_{c\in\delta(x,y)} L(c) \right). \tag{2.3:33}$$

■

The normal sum form of the relative discernibility function $F_D(C)$ indicates set of all D-reducts of the set of attributes C. Each product term of this function represented in this form defines a single D-reduct composed of attributes $c = L^{-1}(\mathsf{c})$ corresponding to Boolean variables in the product term under consideration.

The value reducts in the information system $S = <U, Q, v, f>$ can be obtained using the notion of the value discernibility function for element $x \in U$, which is defined from discernibility matrix $\mathbf{M}(C)$.

Definition 2.3:47 (Value discernibility function, after Skowron and Rauszer 1992)

Each discernibility matrix $\mathbf{M}(C)$ uniquely defines the Boolean value discernibility function for element x, denoted as $F^x(C)$ and given by

$$F^x(C) = \prod_{y\in U:\delta(x,y)\neq\varnothing} \left(\sum_{q\in\delta(x,y)} L(q) \right). \tag{2.3:34}$$

■

The normal sum form of the function $F^x(C)$ indicates set of all value reducts of the set of attributes C for element x. Each product term of the function $F^x(C)$ represented in this form defines a single value reduct for element x. The reduct is composed of the attributes $q = L^{-1}(\mathsf{q})$ corresponding to Boolean variables present in the product term considered.

The relative value reducts for the decision table $T = <U, C, D, v, f>$ are indicated by modified discernibility matrix $\mathbf{M}_D(C)$.

Definition 2.3:48 (Relative value discernibility function, after Skowron and Rauszer 1992)

Each matrix $\mathbf{M}_D(C)$ uniquely defines the Boolean function, called the relative value discernibility function for element x, denoted as $F_D^x(C)$ and given by

$$F_D^x(C) = \prod_{y\in U:\delta(x,y)\neq\varnothing} \left(\sum_{c\in\delta(x,y)} L(c) \right). \tag{2.3:35}$$

■

The normal sum form of the function $F_D^x(C)$ indicates the set of all D-reducts of the set of attributes C for element x. Each product term of the function $F_D^x(C)$ expressed in this form defines a single D-reduct for element x composed of attributes $c = L^{-1}(\mathsf{c})$ corresponding to Boolean variables present in the term under consideration.

The consequence of the assumed indiscernibility relation (5) chosen as the equivalence relation is that QDRSA like CRSA requires discrete values of attributes. This is different from DRSA where notions of reducts and core rely on preference relation, and therefore, this approach does not require discrete attributes. However, at the same time, the assumed preference relation eliminates possibility of the use of the value cores and the value reducts in DRSA. As it will be shown, the advantage of QDRSA over DRSA in some classes of applications lies mainly in natural applicability of these important notions only in QDRSA.

Similarly to DRSA (and contrary to CRSA), QDRSA is dedicated for problems with preference-ordered attributes, however, because QDRSA relies on (5), these attributes need to be of the discrete type. While in some problems it is a clear limitation, in others, namely in such which deal with attributes having inherently discrete nature, the use of classical indiscernibility relation (5) can be advantageous. The illustrative example, concerning the real world application in evolutionary genetics (see section 4.3.3) explains this aspect in more detail. Here, the second limitation of the QDRSA will be given. This limitation is the two-valued domain of the decision attribute $V_d = \{c_0, c_1\}$ where $c_0 < c_1$.

Certainly, the aforementioned constraint excludes QDRSA from being applied in many problems having more complex decisions. However, there is a vast class of applications for which the binary decision is natural and sufficient. In such cases, if the preference-order is in addition naturally assigned to the decision, then application of QDRSA can give better effects that either CRSA (which does not take into consideration the preference order) or DRSA (which resigns from the indiscernibility relation, what, as it will be shown, can lead to sub-optimal solutions).

In general, the types of decision rules obtained in QDRSA are identical to those generated by DRSA. However, because the decision attribute recognizes only two classes and due to relying on equivalence (instead of preference) relation, only two types (out of five possible in DRSA) are generated in QDRSA. These decision rules are of the following types (see Cyran 2009d):

```
if q1 is at least v1 and
   q2 is at least v2 and
       ....          and
   qn is at least vn then
   decision is at least c1
```

and

```
if q1 is at most v1 and
   q2 is at most v2 and
       ....         and
   qn is at most vn then
   decision is at most c0
```

Certainly, if only two classes are recognized the conclusions of the two above types of rules can be changed to *decision is c1* or *decision is c0*, for the first, and the second type, respectively. However, for consistency with DRSA, the full syntax with phrases *at least* and *at most* will be used.

The conditions of the decision rules in QDRSA can be obtained from conditions of the corresponding rules in CRSA by introduction the preference of attribute values to these conditions. Firstly, it requires the change of equalities to phrases like *at least* for the first type conclusion and *at most* for the second type conclusion. Secondly, it requires selection of the minimal set of conditions (considering all decision rules for the given class), since for example the condition *q1 is at least 2* in one rule and *q1 is at least 3* in the other, are subject for dominance relation. This relation is crucial in DRSA. In QDRSA it is also important, but its realm is reduced to the final stage of the information retrieval, as shown above. Therefore in QDRSA, but not in DRSA, the notion of relative value reduct, derivable form the relative value discernibility function $F_D^x(C)$ defined in Definition 47 by equation (35), can be exploited with its full potential.

It is also noteworthy, that not necessarily, the limitation of the types of decision rules to only two aforementioned values, is a drawback. For example, the lack of the fifth type of the decision rules possibly generated by DRSA (see Greco et al. 1999a), is in fact a pure advantage in all problems with binary decision, since senseless in such conditions decision rules of the type

```
if ... then
   decision is at least c0 and at most c1
```

are never generated (contrary to DRSA which in certain situations can generate such rules). Moreover, in the slightly modified syntax, the notation of the two types of rules available in QDRSA is more compact. This syntax uses the notation introduced for QDRSA in Cyran (2009d)

```
if q1 >= v1 and q2 >= v2 and .... and qn >= vn then
   at_least.C1
```

and

```
if q1 <= v1 and q2 <= v2 and .... and qn <= vn then
   at_most.C0
```

which is shorter and therefore it is preferred to be used in QDRSA. In particular it will be used in the illustrative example described in section 4.3.3 to present the advantages of QDRSA over both CRSA and DRSA, in a real application aimed to search the signatures of natural selection operating at molecular level.

2.4. Example: application of considered AI methods

While the practical application of QDRSA is postponed to chapter 4, where search for natural selection is considered, the current section presents the real-world application of the modified by the author indiscernibility relation defined in Definition 2.3:7 by equation (2.3:3) (see section 2.3.2). Remarkably, the application uses also methodology of artificial neural networks described in section 2.2.1 and evolutionary optimization described in section 2.2.2. By utilizing in practical application majority of AI methods described in chapter 2, this application, which serves as an illustrative example, concludes this chapter and supplies the reader with discussion of practical aspects, complementary to theoretical issues considered in sections 2.2 and 2.3.

It is well known, that automatic recognition of images constitutes an important area in the pattern recognition problems based on application of AI methods. In this context, Mrózek and Płonka (1993) were the pioneers in application of rough set models to the image analysis. Studying problem from different perspective, Mait et al. (2003), in a review article, stated that "an examination of recent trends in imaging reveals a movement towards systems that balance processing between optics and electronics". Such systems are designed to perform heavy computations in optical mode, practically contributing no time delays, while post-processing is made in computers, often with the use of AI methods. The foundations of one of such systems have been proposed by Casasent and Song (1985), presenting the design of holographic ring wedge detectors (HRWD), and by George and Wang (1994), who combined commercially available ring wedge-detector (RWD) and neural network in a one complete image recognition system.

Despite the completeness of the solution their system was of little practical importance, since commercially available RWD was very expensive and moreover, it could not be adapted to a particular problem. Casasent's HRWD, originally named by him as a computer generated hologram (CGH) had a lot of advantages over commercial RWD, most important being: much lower cost and adaptability. According to optical characteristics the HRWD belongs to a wider class of grating based diffractive optical variable devices (DOVDs) (Cyran et al. 2001c), which could be relatively easy obtained from the computer generated masks, and which can be used for sampling the Fraunhofer diffraction patterns.

The pioneering works proposing the method of optimization of HRWD masks to a given application have been published by Jaroszewicz et al. (2000) and by Cyran and Mrózek (2001). Mentioned method was successfully applied to a MLP-based system, in a recognition of the type of subsurface stress in materials with embedded optical fiber (Cyran et al. 2001b, 2002). The examples of application of the RWD-based feature extraction together with MLP-based classification module include systems designed by Podeszwa et al. (2003) devoted for

the monitoring of the engine condition, and by Jaroszewicz et al. (2002) dedicated for airplane engines.

Some other notable examples of applications of ring-wedge detectors and neural network systems, include works of Ganotra et al. (2003), and Benfanger and George (1999), concerning fingerprint recognition, face recognition (Ganotra et al. 2002), or image quality assessment (Berfanger and George 2000). The ring-wedge detector has been also used, as a light scatter detector, in a classification of airbone particles performed by Kaye et al. (2000) and accurate characterization of particles or defects, present on or under the surface, useful in fabrication of integrated circuits, as presented by Nebeker and Hirleman (2000).

The purely optical version of HRWD-MLP recognition system was considered by Cyran and Jaroszewicz (2001), however, such system is limited by the development of optical implementation of neural networks. Simplified, to rings only, version of the device is reported by Fares et al. (2000) to be applied in a rotation invariant recognition of letters. With all these applications, no wonder that Mait et al. (2003) concluded:" few attempts have been made to design detectors with much consideration for the optics. A notable exception is ring-wedge detector designed for use in the Fourier plane of a coherent optical processor."

Obviously, MLP (or more generally any type of NN) is not the only classifier which could be applied for classification of patterns occurring in a feature space generated by HRWD. Moreover, the first version of optimization procedure favored the rough set based classifiers, due to identical (and therefore fully compatible) discrete nature of knowledge representation in the theory of rough sets applied both to HRWD optimization and to subsequent rough set based classification. The application of general ideas of obtaining such rough classifier was presented by Cyran and Jaroszewicz (2000) and fast rough classifier implemented as PAL 26V12 element was considered and designed by Cyran (2003). Despite of inherent compatibility between optimization procedure and the classifier, the system remained sub-optimal, because features extracted from HRWD generate continuous space, subject to unnatural discretization required by both: rough set based optimization and classifier.

Mentioned problems led to the idea, that in order to obtain the enhanced optimization method, the discretization required by classical indiscernibility relation in rough set theory, should be eliminated in such a way, which does not require the resignation from the equivalence relation in a favor of some weaker form (like tolerance relation, for example). It was achieved by such modification of the indiscernibility relation, which allows for natural processing of the real valued attributes (this problem is considered in detail in the section 2.3.2). The current section start with optical foundations of the recognition system considered, and it is followed by experimental results obtained after application of the enhanced optimization methodology.

Remarkably, the experimental application of the modified indiscernibility relation presented in the section 2.3.2, to the system considered, improved the results of evolutionary optimization of holographic RWD and equivalently, enhanced the optimization of the HRWD generated feature space, dedicated for real-valued classifiers. It also gave theoretical basis for the design of two-way, neural network-rough set based classification system (Cyran 2005b).

As it has been mentioned, presented below system belongs to a class of fast hybrid opto-electronic pattern recognizers. Since, feature extraction subsystem is processing the information optically, let us start a description of such feature extractor by giving a physical basis, required to understand the properties of feature vectors generated by this subsystem. This introductory material will be followed by the description of enhanced author's method of HRWD optimization and experimental results of the usage of this optimization. This illustrative section is completed with the description of probabilistic neural network (PNN) based classifier and experimental results of the application of it into the Fraunhofer pattern recognition.

Consider homogeneous and isotropic medium which is free of charge ($\rho = 0$) and currents ($j = 0$). In such medium (see Cyran 2008b), Maxwell equations result in the absence of charges and currents, in a wave equation

$$\nabla^2 \mathbf{G} - \varepsilon'\mu' \frac{\partial^2 \mathbf{G}}{\partial t^2} = 0 \tag{2.4:1}$$

where **G** denotes electric (**E**) or magnetic (**H**) field, and a product $\varepsilon'\mu'$ is the reciprocal of squared velocity of a wave in a medium. Application of this equation to a space with obstacles like apertures or diaphragms should result in equations describing the diffraction of the light at these obstacles. However the solution is very complicated for special cases and impossible for the general case. Therefore the simplification should be used which assumes a scalar field u instead of vector field **G**. In such a case the information about the light polarization is lost. For such scalar field it holds that (see Cyran 2008b)

$$\nabla^2 u - \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} = 0. \tag{2.4:2}$$

Simplified in this way theory, called the scalar Kirchhoff's theory, describes the diffraction of the light at various obstacles. According to this theory, scalar complex amplitude $u_0(P)$ of a light oscillation, caused by the diffraction, is given in a point of observation P by the Kirchhoff's integral (Piekara 1976)

$$u_0(P) = \frac{1}{4\pi} \int_{\Sigma} \left[\frac{e^{ikr}}{r} \frac{du_0}{dn} - u_0 \frac{d}{dn}\left(\frac{e^{ikr}}{r} \right) \right] d\Sigma \tag{2.4:3}$$

where Σ denotes closed surface with point *P* and without the light source, *n* is an external normal to the surface Σ, $k = 2\pi/\lambda$ is a propagation constant, u_0 denotes scalar amplitude on a surface Σ, and *r* is the distance between any point covered inside surface Σ to the observation point *P*. Formula (3) states that amplitude u_0 in point *P* does not depend on the state of oscillations in the whole area surrounding this point (what would result from Huygens theory) but, depends only on state of oscillations on a surface Σ. All other oscillations inside this surface are canceling each other. Application of Kirchhoff's theorem to a diffraction on a flat diaphragm with aperture of any shape and size gives the integral stretched only on a surface Σ_A covering the aperture. Such integral can be transformed to (Piekara 1976):

$$u_0(P) = -\frac{ik}{4\pi}\int_{\Sigma_A} u_0(1+\cos\theta)\frac{e^{ikr}}{r}\,d\Sigma_A \tag{2.4:4}$$

where θ denotes an angle between radius *r* from any point of aperture to point of observation, and the internal normal of the aperture.

Since any transparent image is, in fact, a collection of diaphragms and apertures of various shapes and sizes, therefore such image, when illuminated by coherent light, generates the diffraction pattern, described in scalar approximation by the Kirchhoff's integral (3). Let coordinates of any point *A*, in an image plane, are denoted by (*x*, *y*), and let an amplitude of light oscillation in this point, be *v* (*x*, *y*). Furthermore, let coordinates (ξ, η) of an observation point *P* be chosen as (Cyran 2008b):

$$\xi = \frac{2\pi}{\lambda}\sin\theta, \qquad \eta = \frac{2\pi}{\lambda}\sin\varphi \tag{2.4:5}$$

where: λ denotes the length of the light wave, whereas θ and φ are angles between the radius from the point of observation *P* to point *A*, and planes (*x*, *z*) and (*y*, *z*), respectively.

These planes are two planes of such coordinate system (*x*, *y*, *z*), whose axes *x* and *y* are in the image plane, and axis *z* is perpendicular to the image plane (it is called optical axis). Let coordinate system (*x*', *y*') be the system with the beginning at point *P* and such that its plane (*x*', *y*') is parallel to the plane of the coordinate system (*x*, *y*). It is worth to notice, that coordinates of one particular point in the observation system (ξ, η) correspond to coordinates of all points *P* of the system (*x*', *y*'), such that the angles between axis *z* and a line connecting these points with some points *A* of the plane (*x*, *y*), are θ and φ, respectively.

In other words, all radii *AP*, connecting points *A* of the plane (*x*, *y*) and points *P* of the plane (*x'*, *y'*), which are parallel to each other, are represented in a system (ξ, η) by one point. Such transformation of the coordinate systems is physically obtained in the back focal plane of the lens, placed perpendicularly to the optical axis *z*. In this case, all parallel radii represent parallel light beams, diffracted on the image (see Fig. 1) and focused in the same point in a

focal plane. Moreover, the integral (3), when expressed in a coordinate system (ξ, η), can be transformed to (Piekara 1976):

$$u_0(\xi,\eta)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}v(x,y)e^{-i(\xi x+\eta y)}dxdy. \tag{2.4:6}$$

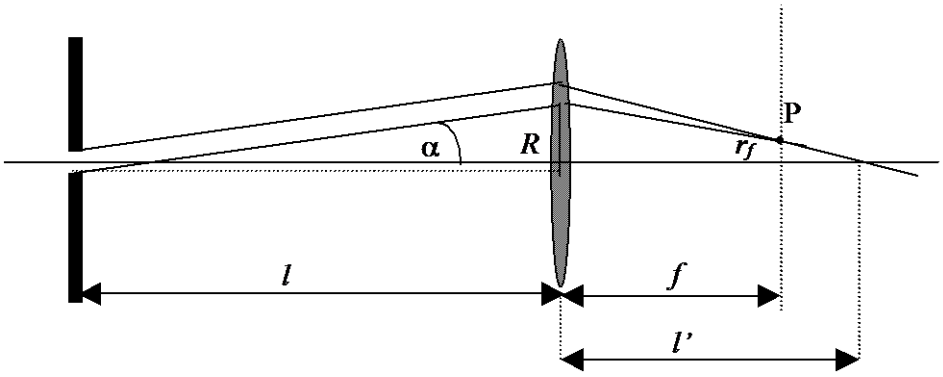


Fig. 2.4:1. The operation of the spherical lens (after Cyran 2008b)
Rys. 2.4:1. Działanie soczewki sferycznej (na podstawie Cyran 2008b)

Geometrical relationships in Fig. 1 reveal that

$$r_f = R\frac{l'-f}{l'}. \tag{2.4:7}$$

On the other hand the operation of the lens is given by

$$\frac{1}{f}=\frac{1}{l}+\frac{1}{l'}. \tag{2.4:8}$$

Letting equation (8) to (7), after elementary algebra, one obtains

$$\frac{R}{l}=\frac{r_f}{f}. \tag{2.4:9}$$

Since angles θ and φ (corresponding to angle α in Fig. 1, in a plane (x, z) and (y, z), respectively) are small, therefore equations (5), having in mind (9), can be rewritten as (Cyran 2008b)

$$\xi=\frac{2\pi}{\lambda}\frac{x_f}{f}, \qquad \eta=\frac{2\pi}{\lambda}\frac{y_f}{f} \tag{2.4:10}$$

where x_f and y_f denote Cartesian coordinates in a focal plane of the lens. Equation (6) expressed in these coordinates can be written as (Cyran 2008b)

$$u_0(x_f,y_f)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}v(x,y)e^{-i2\pi\left(\frac{x_f}{\lambda f}x+\frac{y_f}{\lambda f}y\right)}dxdy. \tag{2.4:11}$$

Finally, setting new coordinates (u, v) as

$$u = \frac{x_f}{\lambda f}, \qquad v = \frac{y_f}{\lambda f} \tag{2.4:12}$$

the equation (see Cyran 2008b)

$$u_0(u,v) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} v(x,y) e^{-i2\pi(ux+vy)} dxdy \tag{2.4:13}$$

can be derived, which is (up to the constant factor k) a Fourier integral. This is essentially the Fraunhofer approximation of Kirchhoff's integral, and is also referred to as a Fraunhofer diffraction pattern (Kreis 1996). The complex amplitude of the Fraunhofer diffraction pattern obtained in a back focal plane of the lens is therefore a Fourier transform of the complex amplitude from the image plane

$$u_0(u,v) = k\Im\{v(x,y)\}. \tag{2.4:14}$$

This fact is very often used in a design of hybrid systems for recognition of images in a spatial frequency domain. One prominent example is the system with a feature extractor built as a HRWD placed in a back focal plane of the lens. The HRWD itself consists of two parts: a part composed of rings R_i and a part containing wedges W_j.

In a holographic version of ring-wedge detector, each of elements R_i or W_j is covered with a grating of particular spatial frequency and orientation, so that the light, passing through the given region, is diffracted and focused by some other lens, at certain cell of array of photodetectors. The photodetector, in turn, integrates the intensity of the light and generates one feature used in classification. Since two-dimensional Fourier transform satisfies properties:

$$\Im\{f(x-a, y-b)\} = F(u,v)\exp[-2\pi i(au+bv)]. \tag{2.4:15}$$

$$F(-u,-v) = F^{\#}(u,v). \tag{2.4:16}$$

stating, that power spectrum of the input signal is shift invariant and symmetrical with respect to center of the spatial frequency coordinate system, and that all information about the light intensity in the Fourier plane is covered in every half-plane with the edge crossing the center of the optical system, therefore each half-circle of the HRWD samples full and shift invariant information describing the input image.

Moreover, the power spectrum satisfies formulae:

$$\begin{aligned} |\Im\{f(x\cos\alpha - y\sin\alpha, x\sin\alpha + y\cos\alpha)\}|^2 = \\ = |F(u\cos\alpha - v\sin\alpha, u\sin\alpha + v\cos\alpha)|^2 \end{aligned}. \tag{2.4:17}$$

and

$$|\Im\{f\{ax, ay)\}|^2 = |\frac{1}{a^2} F(\frac{u}{a}, \frac{v}{a})|^2. \tag{2.4:18}$$

concerning rotation and rescaling of Fourier image. According to these formulae wedges integrating light intensity generate scale invariant elements of feature vector. Similarly, rings generate rotation invariant information.

To avoid the superposition of first order beam with higher order beams, diffracted by HRWD, the distance d_{ij} between two lines of grating covering its regions must satisfy the equation (see Cyran 2000)

$$d_{ij} = \frac{\lambda\, f_L}{h_{i,1}} \cos\theta_{ij} = \frac{\lambda\, f_L}{H - \left(i - \frac{1}{2}\right)s} \cos\theta_{ij}. \tag{2.4:19}$$

where an angle θ_{ij} which they form with horizontal axis of the HRWD is given by

$$\theta_{ij} = \operatorname{arctg}\left[\frac{\left(j - \frac{1}{2}\right)s}{H - \left(i - \frac{1}{2}\right)s}\right]. \tag{2.4:20}$$

In the above formulae indices i and j correspond to the row and the column in the photodetector rectangular array. Designations H and S are graphically explained in Fig. 2. Features obtained by HRWD, after being converted by an array of photodetectors into electronic signals, are used as the input by the AI-based classifier.

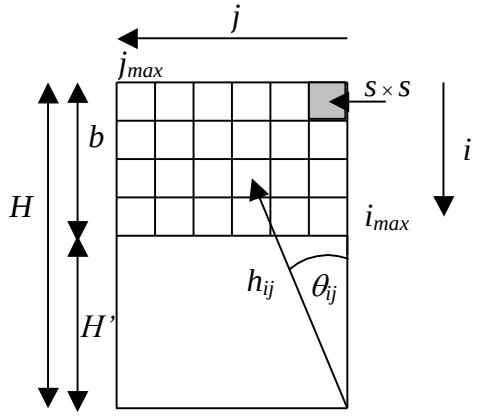


Fig. 2.4:2. Array of photodetectors converting the light intensities into the electronic features (after Cyran and Mrózek 2001)
Rys. 2.4:2. Matryca fotodetektorów zmieniających intensywność światła na cechy elektroniczne (na podstawie Cyran and Mrózek 2001)

The system considered above can be used for the recognition of images invariant with respect to translation, rotation and size, based on the properties of the Fourier transform and the way of sampling the Fraunhofer diffraction pattern by the HRWD. Standard HRWD based feature extractor can be optimized to obtain even better recognition properties of the

system. To perform any optimization one needs the objective function and the method of search in a space of solutions. These two problems are discussed wider below.

Let ordered 5-tuple $T = <U, C, \{d\}, v, f>$ be the decision table obtained from the information system $S = <U, Q, v, f>$ by a decomposition of the set of attributes Q into two mutually disjoint sets: the set of conditional attributes C and the set $\{d\}$composed of one decision attribute d. Let each conditional attribute $c \in C$ be one feature obtained from HRWD, and let decision attribute d be the number of the class to be recognized. Obviously the domain of any of such conditional attributes is $\Re$ and the domain of decision attribute d is a subset of first natural numbers, with cardinality equal to the number of recognized classes.

Furthermore, let $\mathbf{D} = \{ [x_n]_{I_0(\{d\})} : x_n \in U\}$ be the family of such sets of images where each set contains all images belonging to the same class. Observe that the classical form of the indiscernibility relation I_0 is used in this definition, due to discrete nature of the domain of decision attribute d.

Based on the results of discussion given by Cyran and Mrozek (2001), it follows that the rough set based coefficient, called quality of approximation of family **D** by conditional attributes belonging to C, and denoted by $\gamma_C(\mathbf{D})$, is a good objective function in the optimization of feature extractor in problems with multimodal distribution of classes in a feature space. This is so, because this coefficient indicates the level of determinism of the decision table, what in turn, is relevant for the classification.

On the other hand, based on discussion given in section 2.3.2, in the case of real valued attributes C, the preferred form of indiscernibility relation, being so crucial for rough set theory in general (and therefore for the computation of $\gamma_C(\mathbf{D})$ objective in particular), is the form defined by (2.3:3). Therefore the optimization with the objective function $\gamma_C(\mathbf{D})$ computed with respect to classical form of indiscernibility relation for real valued attributes C given in (2.3:2) produces sub-optimal solutions. This drawback can be eliminated if modified version proposed in (2.3:3) is used instead of classical form defined in (2.3:2).

However, the generalized form (2.3:3) requires the definition of some structure in a set of conditional attributes. This is task dependent, and in the case considered, the architecture of the feature extractor having different properties of wedges and rings, defines natural structure, as a family $\mathbf{C} = \{C_R, C_W\}$, composed of two sets: a set of attributes corresponding to rings C_R, and a set of attributes corresponding to wedges C_W. With this structure introduced into set of conditional attributes, the coefficient $\gamma_C(\mathbf{D})$ computed with respect to modified indiscernibility relation (2.3:3), is en enhanced objective function for optimization of the HRWD.

Since, the defined above enhanced objective function is not differentiable, gradient-based search method should be excluded. However the HRWD can be optimized in a framework of

slightly modified evolutionary algorithm (for details of evolutionary computation see section 2.2.2), as presented in pseudo-code below (see Cyran and Niedziela 2009):

```
POPULATION ← Initialize;
t ← 1; Evaluate (Q); ξ ← 2^Q;
do for x in POPULATION
  do for i = 1 to card (U)
    C_x[i] ← χ (image_i ); d_x[i] ← C_j ;
  od;
  I_1 ← Evaluate (Clusterize(C));
  F_x ← Evaluate (γ_C (D*));
od;
do while (ξ ≥ NumOfClasses) and (t < MaxGenNum)
  FOUND ← FALSE;
  POPULATION ← Select (POPULATION);
  POPULATION ← Recombine (POPULATION);
  POPULATION ← Mutate (POPULATION);
  POPULATION ← Repair (POPULATION);
  do for x in POPULATION
    do for i = 1 to card (U)
      C_x[i] ← χ (image_i ); d_x[i] ← C_j ;
    od;
    I_1 ← Evaluate (Clusterize(C));
    F_x ← Evaluate (γ_C (D*));
    if F_x = MaxValue then
      FOUND ← TRUE;
      x_opt ← x;
    fi;
  od;
  if FOUND then
    ξ ← ξ / 2;
  fi;
  t ←  t + 1;
od;
```

In the above algorithm t is the generation number, $\mathbf{x}$ is the chromosome (representing the HRWD) in population POPULATION and $\mathbf{x}_{opt}$ is the chromosome representing genotype of the optimum HRWD. $C_x[i]$ are discrete conditions of decision rule i generated by HRWD for image $image_i$. Similarly, $d_x[i]$ denotes the decision attribute of mentioned decision rule and C_j is the abstract class the image $image_i$ belongs to.

As genetic operations, classical one point recombination and uniform mutation, have been used. The selection was proportional, however in the elitist model, propagating the best solution from generation to generation, with probability 1. To retain the solutions in a space of allowed by phenotype constraints limits, the repair algorithm was applied, after genetic operations.

The algorithm has two flow control parameters: *MaxGenNum* (specifying maximum number of epochs for evolution) and *MaxValue,* indicating the maximum required value of the objective function. Normally *MaxValue* should be set to 1, to obtain fully consistent decision table, but sometimes this could be too strong demand to fulfill – then one should reduce this parameter.

This algorithm is very similar to that, applied in the case of the objective function, calculated from the classical definition of the discernibility relation. The difference is in the meaning of ξ parameter. When the classical indiscernibility relation is used, ξ is a discretization factor, required by the rough set theory. On the other hand, when modified version of indiscernibility relation defined by (2.3:3) is applied, ξ is the number of clusters in a clustering procedure. This change influences the initial value of ξ and the termination of presented program. The initial value of ξ for modified indiscernibility relation is calculated as 2^Q for such minimum Q, for which $\xi \geq Card\,(U)$.

The program is terminated after achieving the maximum value of $\gamma_C(D^*) = MaxValue$, for $\xi = NumOfClasses$ (*NumOfClasses* denotes the number of classes to be recognized), as opposed to classical version (see Cyran and Mrózek 2001), terminating when $\gamma_C(D^*) = MaxValue$, for $\xi = 2$. Another difference is, that in the above algorithm, the function χ denotes the feature extraction, while in the classical version it denoted the feature extraction with discretization, so the clustering has to be invoked explicitly. As the result of operation of the algorithm, the parameters describing optimized HRWD are obtained (they are encoded in chromosome $\mathbf{x}_{opt}$). The results of this algorithm, in a form of a time course of the objective function, are presented in Fig. 3 in liner and in Fig. 4 in the logarithmic scale.

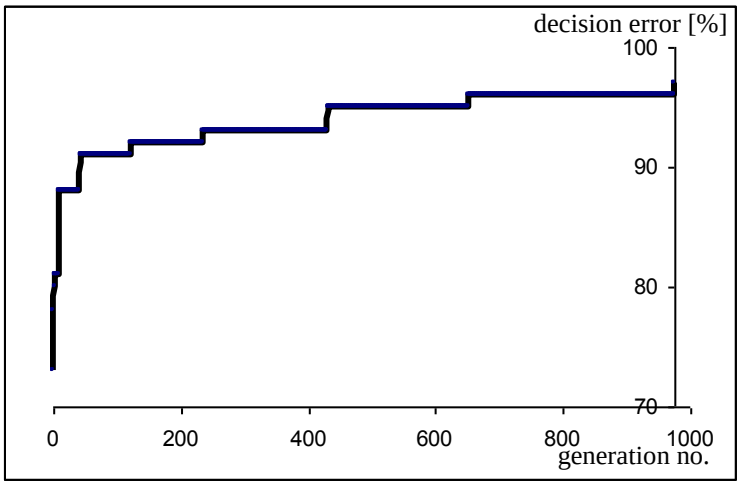


Fig. 2.4:3. Process of evolutionary optimization of HRWD for discretization factor ξ = 16 in linear scale (after Cyran 2008b)

Rys. 2.4:3. Proces ewolucyjnej optymalizacji HRWD dla współczynnika dyskretyzacji ξ = 16 w skali liniowej (na podstawie Cyran 2008b)

The two graphs given in Fig. 3 and Fig 4. present the fitness of $\mathbf{x}_{opt}$ expressed in percents. As defined above, the family of conditional attributes $\mathbf{C} = \{C_R, C_W\}$, where C_R denotes attributes generated by rings and C_R denotes attributes generated by wedges. The maximum value of fitness 97%, having the meaning of $\gamma_C(D^*) = 0.97$, was obtained in 976 generation for population composed of 50 individuals.

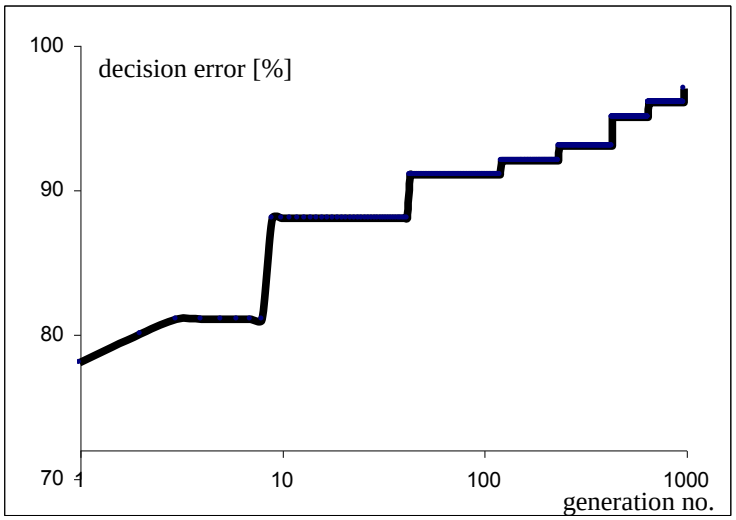


Fig. 2.4:4. Process of evolutionary optimization of HRWD for discretization factor ξ = 16. The course uses logarithmic horizontal scale on axis indicating the number of generations (after Cyran 2008b)

Rys. 2.4:4. Proces ewolucyjnej optymalizacji HRWD dla współczynnika dyskretyzacji ξ = 16. Wykres wykorzystuje skalę logarytmiczną na osi poziomej wskazującej ilość pokoleń (na podstawie Cyran 2008b)

The computer generated mask of optimal HRWD, encoded by $\mathbf{x}_{opt}$ is presented in Fig. 5b. In Fig 5a the mask, optimized with classical indiscernibility relation, is given for comparison.

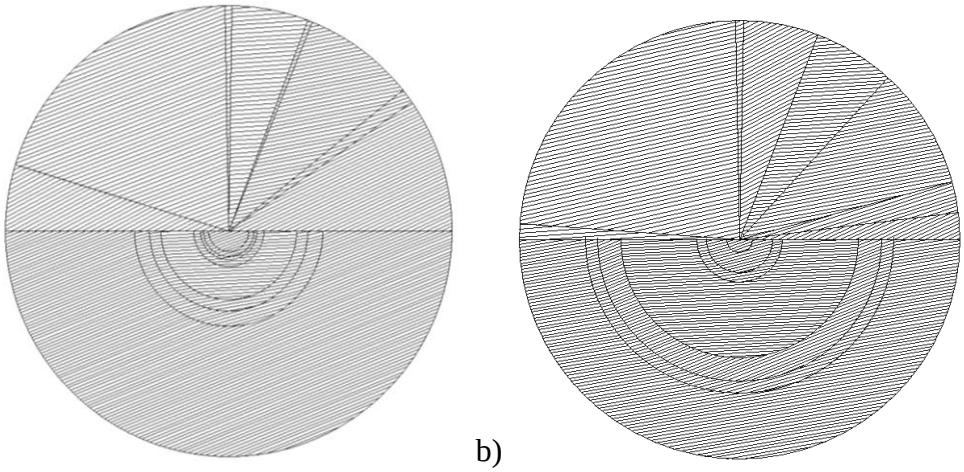


Fig. 2.4:5. The computer generated mask of HRWD optimized with a) classical indiscernibility relation, b) modified indiscernibility relation (after Cyran and Niedziela 2009)

Rys. 2.4:5. Komputerowo generowane maski HRWD optymalizowanego z a) klasyczną relacją nierozróżnialności, b) zmodyfikowaną relacją nierozróżnialności (na podstawie Cyran i Niedziela 2009)

These masks are designed for a system with a coherent light wave length $\lambda = 635$ nm, emitted by laser diode and for a lens L with a focal length $f_L = 1$ m. In order to keep the

resolution capability of the system, the diameter of the HRWD in a Fourier plane should be equal to the diameter of the Airy disc, which is given by the equation: $s_{\mathrm{HRWD}} = 4 \times 1.22 \times \lambda \times f_{L_1} / s_{min} = 2.07$ mm, if the assumed minimum size of recognizable objects $s_{min} = 1.5$ mm. Assuming also the rectangular array of photodetectors of the size $s = 5$ mm (see Fig. 2 for the exact meaning of designation s and subsequent symbols), forming four rows ($i = 1,...4$) and four columns ($j = 1,...,4$), and setting $H = 50$ mm it is possible to obtain the values of angles θ_{ij} given by (20) as presented in Table 1.

Table 2.4:1

The values of angles θ_{ij} (expressed in degrees) defining the HRWD gratings (after Cyran 2008b)

4	3	2	1	$\leftarrow j,\ i\downarrow$
20.22	14.74	8.97	3.01	1
22.38	16.39	10.01	3.37	2
25.02	18.43	11.31	3.81	3
28.30	21.04	12.99	4.40	4

Similar results for the distances d_{ij} given by (19) are presented in Table 2.

Table 2.4:2

Distances d_{ij} between striae [μm] (after Cyran 2008b)

4	3	2	1	$\leftarrow j,\ i\downarrow$
12.54	12.93	13.20	13.35	1
13.82	14.33	14.71	14.92	2
15.34	16.06	16.60	16.90	3
17.20	18.24	19.04	19.48	4

Since the software developed by the author for generating HRWD masks has been designed in such a way, that the distances d_{ij} are given in units equal to a one-tenth of a percent of the radius of HRWD, therefore in Table 3 for $R_{\mathrm{HRWD}} = s_{\mathrm{HRWD}} / 2 = 1.035$ mm, the proper values expressed in these units are presented.

Table 2.4:3

Distances d_{ij} between striae, in units used by software generating HRWD masks (after Cyran 2008b)

4	3	2	1	$\leftarrow j,\ i\downarrow$
12.14	12.52	12.78	12.92	1
13.38	13.88	14.24	14.44	2
14.86	15.55	16.08	16.36	3
16.65	17.65	18.43	18.86	4

The classification, i.e., the transformation δ, from a space of vectors *V*, to a set of classes *C*, becomes in a supervised version a mapping known for *P* examples (v_i, $\omega_{j(i)}$), where v_i ($i = 1, ..., P$) are feature vectors, and $\omega_{j(i)}$ ($j(i) = 1, ..., M$) are the associated classes. For the experimental verification of applicability of the proposed indiscernibility relation modification, there was built the optimal classifier in probabilistic uncertainty model. It is the mapping, minimizing the Bayesian risk *R*, given by (Jutten 1997)

$$R = \sum_{i=1}^{M} \sum_{j=1}^{M} L_{ij} P_j \int \cdots \int_{\mathbf{v} \in D_i} p\left(\mathbf{v} \middle| \omega_j\right) d\mathbf{v} . \tag{2.4:21}$$

where P_j is the prior probability of the class ω_j, D_i is the region in feature space, in which each point is assigned to the class ω_i, and L_{ij} is the cost of decision: ω_i while ω_j is true. Assuming equal loss, associated with any bad classification, i.e., $L_{ij} = 1 - \delta_{ij}$, where δ_{ij} is the Kronecker symbol, and rearranging equation (21) for $j = i$, and $j \neq i$, it follows that (Jutten 1997)

$$R = \sum_{i=1}^{M} \sum_{j \neq i} P_j \int \cdots \int_{\mathbf{v} \in D_i} p\left(\mathbf{v} \middle| \omega_j\right) d\mathbf{v} = \sum_{i=1}^{M} \int \cdots \int_{D_i} \left(\sum_{j \neq i} P_j p\left(\mathbf{v} \middle| \omega_j\right) \right) d\mathbf{v} . \tag{2.4:22}$$

Since each integrand I_i is positive, therefore the risk *R* is minimized if, and only if, the feature vector **v** is assign to such class $\omega_{k(\mathbf{v})}$, that (Cyran and Niedziela 2009)

$$k(\mathbf{v}) = \arg \min_{1 \le i \le M} \sum_{j \neq i} P_j p\left(\mathbf{v} \middle| \omega_j\right) d\mathbf{v} . \tag{2.4:23}$$

or, what is equivalent:

$$k(\mathbf{v}) = \arg \max_{1 \le i \le M} \left(P_i p\left(\mathbf{v} \middle| \omega_i\right) d\mathbf{v} \right). \tag{2.4:24}$$

Moreover, if Bayesian rule is applied, equation (24) can be transformed to

$$k(\mathbf{v}) = \arg \max_{1 \le i \le M} \left(p\left(\omega_i \middle| \mathbf{v}\right) d\mathbf{v} \right). \tag{2.4:25}$$

which is well known maximum posterior probability principle.

Equation (25) defines the decision rule used in any statistical classifier, whereas equation (24) can be especially easy implemented as a PNN, described in section 2.2.1. In the design considered, the input layer is composed of *N* elements to process *N*-dimensional feature vectors generated by HRWD ($N = N_R + N_W$) (Fig. 6). The pattern layer consists of *M* pools of pattern neurons, associated with *M* classes of intermodal interference to be recognized. The *j*th pool in the pattern layer is built up of S_j = card (V_j) nodes. In that layer, the RBF neurons with Gaussian transfer function have been used for implementation of the kernel function (2.2:60).

Then, the width of the kernel function is simply a standard deviation σ of the Gaussian bell. Additionally, when using such networks as classifiers, formally, there is a need to

multiply the output values by prior probabilities P_j, in order to be able to apply the decision rule described by (24). However in the case considered, all priors are equal and therefore, the values given by (24) can be obtained directly on the PNN outputs defined by (2.2:60).

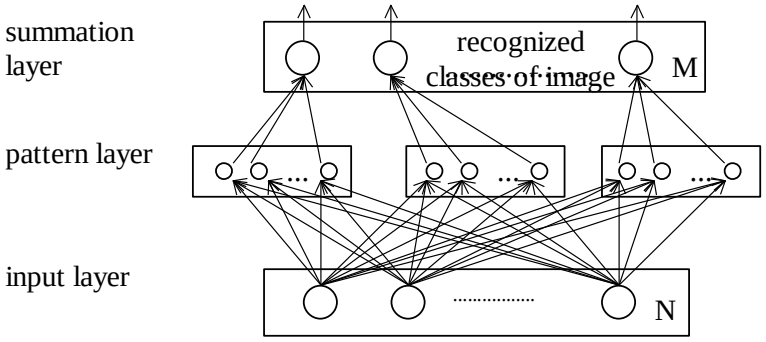


Fig. 2.4:6. Probabilistic neural network classifying features obtained from optimized HRWD (after Cyran and Niedziela 2009)
Rys. 2.4:6. Sieć neuronowa probabilistyczna klasyfikująca cechy otrzymane z optymalizaowanego HRWD(na podstawie Cyran i Niedziela 2009)

The verification of the recognition abilities was performed by a classification of the speckle structure images, obtained from the output of the optical fiber. The experiments were conducted for a set of 128 images of speckle patterns generated by intermodal interference occurring in optical fiber and belonging to eight classes taken in 16 sessions S_l (l = 1, ..., 16). The Fraunhofer diffraction patterns of the input images were obtained by calculating the intensity patterns from the discrete Fourier transform equivalent to (13).

The training set consisted of 120 images, taken out in 15 sessions, and the testing set contained 8 images, belonging to the different classes, representing one session S_l. The process of training and testing was performed 16 times, according to *delete*-8 jackknife method, i.e., for each iteration, another session composed of 8 images was used for the testing set, and all but one sessions were used for the training set. That gave the basis for reliable cross-validation with still reasonable number of images used for training, and the reasonable computational time. This time was eight times shorter, as compared to the classical leave-one-out method, which is practically equivalent to *delete*–1 jackknife method (the only difference is the resubstitution error of a prediction model but this problem will not be addressed here).

The jackknife method was used for cross validation of PNN results, because of the unbiased estimation of the true error in the probabilistic classification model (contrary to the underestimated error - however having smaller variance – obtained by the Bootstrap method) (Twomey and Smith 1998, Azuaje 2003). Therefore, the choice of the *delete*-8 jackknife method, was a sort of tradeoff between the accuracy (standard deviation of estimated

normalized decision error was 0.012), the unbiased estimate of the error, and the computational effort. The results of such testing of the PNN applied to the classification of the images in the feature space obtained from the standard, optimized, and the optimized with modified indiscernibility relation HRWDs, are presented in Table 4.

These results were obtained with the PNN classifier having Gaussian radial function with standard deviation $\sigma = 0.125$. In the last column of Table 4 the improvement is computed with respect to the standard HRWD (the first value) and with respect to HRWD optimized with standard indiscernibility relation (the value in a parentheses).

Table 2.4:4
Results of testing the classification abilities of the HRWD-PNN system (after Cyran 2008b)

	Correct decisions [%]	Normalized Decision error [%]	Improvement [%]
Standard HRWD	84.4	1.95	0.0 (-25.0)
HRWD optimized with standard indiscernibility relation	87.5	1.56	20.0 (0.0)
HRWD optimized with modified indiscernibility relation	88.3	1.46	25.1 (6.4)

More detailed results of all jackknife tests are presented in Table 5, Fig. 7 and Fig. 8. In Table 5 bold font is used for the results differing between optimization with standard and modified version of indiscernibility relation. Bold underlined results indicate improvement when modified relation is used instead of classical. Bold results without underlining indicate the opposite.

In Fig. 7, the horizontal axis represents the number of the test, while the vertical axis is a cumulative (*i.e.* cumulating from the first test to test with number given on the horizontal axis) number of bad decisions. Observe, that starting from test number 9 to the end, the cumulative number of bad decisions is better for optimization of HRWD performed with modified indiscernibility relation, as compared to optimization with classical version of this relation. The opposite situation is only for test number 6, while in tests number 1 to 5 both procedures performed equally good, outperforming results obtained with application of standard, not optimized HRWD.

Table 2.4:5

Detailed results of PNN testing for the tests number 1 to 16 (after 2008b)

NUMBER OF TEST SESSION:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	NUMBER OF BAD DECISIONS															
Standard HRWD	1	2	2	1	2	0	1	0	1	1	4	0	0	1	0	4
HRWD optimized with standard indiscernibility relation	1	1	3	0	1	0	2	0	2	1	1	0	0	1	1	2
HRWD optimized with modified indiscernibility relation	1	1	3	0	1	**1**	**1**	0	**1**	1	1	0	0	1	1	2

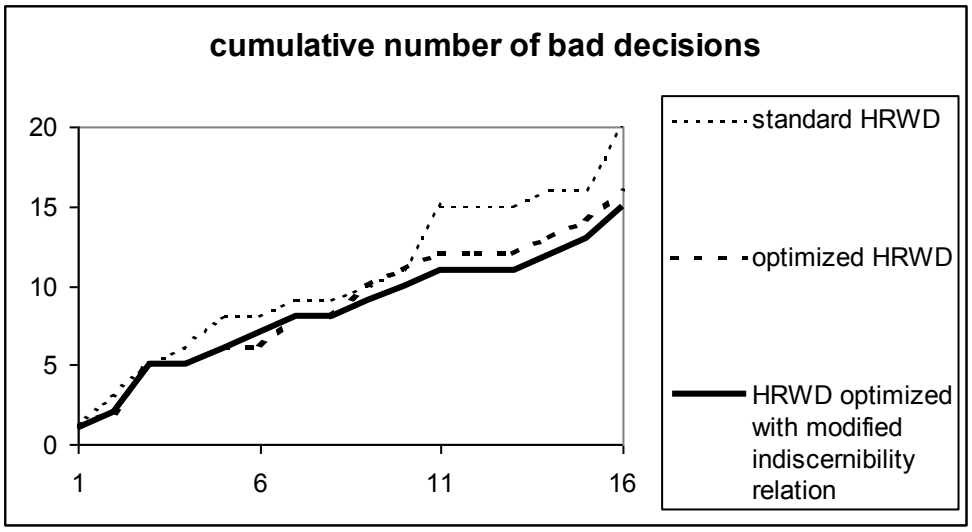


Fig. 2.4:7. Graphical representation of the cumulative results of testing in the HRWD-PNN system (after Cyran 2008b)

Rys. 2.4:7. Graficzna reprezentacja skumulowanych rezultatów testów w systemie HRWD-PNN (na podstawie Cyran 2008b)

In Fig. 8, the horizontal axis represents the number of the test, while the vertical axis is a normalized decision error averaged over tests from the first to given, represented by the value of horizontal axis. Observe, that for averaging over more than 8 tests, the results for recognition with HRWD optimized with modified indiscernibility relation are outperforming both: results for HRWD optimized with classical version of indiscernibility relation and results for standard HRWD.

The normalized decision errors, ranging from 1.5 to 2 percent, indicate good overall recognition abilities of the system. The 20% reduction of this error is obtained by

optimization of HRWD with classical indiscernibility relation. Further 6% error reduction, is caused solely by the modification of the indiscernibility relation, according to (2.3:3). In order to understand the scale of this improvement, not looking too impressive at first glance, one should refer to Fig. 3 and take into consideration, that this additional 6% error reduction is obtained over an already optimized solution.

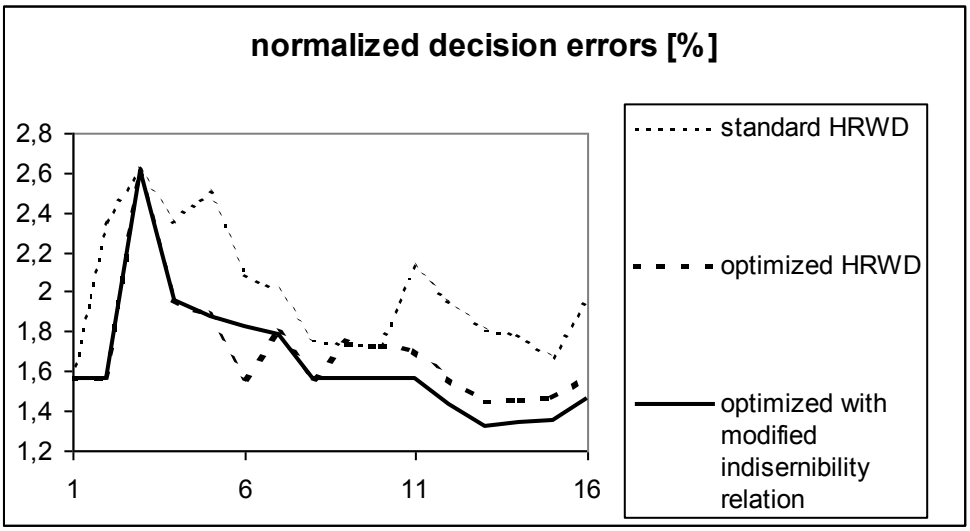


Fig. 2.4:8. Graphical representation of the normalized decision error during testing in the HRWD-PNN system (after Cyran 2008b)
Rys. 2.4:8. Graficzna reprezentacja znormalizowanego błędu decyzji dla testów w systemie HRWD-PNN (na podstawie Cyran 2008b)

The level of difficulty can be grasped observing that, on average, the increase of the objective function is well mimicked by a straight line, if a generation number axis is drawn in a log scale (see Fig. 4). This means, that the growth of the objective is, on average, well approximated by a logarithmic function of the generation number. It experimentally reflects a well known fact, stating that, the better current solution is, the harder is to optimize it further (more specifically: harder, means that it requires more generations in evolutionary process).

2.5. Conclusions

In this chapter the field of artificial intelligence was discussed in context of the author's scientific work in this area. After giving the general introduction to different approaches within this domain such as strong AI or weak AI, a more detailed part follows, which deals with biologically inspired AI-methods (artificial neural networks, and evolutionary computation) in section 2.2 and rough set based approaches (section 2.3). Such choice is influenced by interests of the author as well as his research which is focused in these three domains. Therefore, many important for AI issues have not been discussed. In particular, the

recent trends in hardware which is expected to implement advanced AI-based methods were not considered. The monograph as a whole does not discuss advances in hardware technologies, however it is worth to notice that perhaps the emergence of many phenomena considered by strong AI will be possible only in hardware technologies, which are radically different from classical. Such technologies as quantum systems (see for example Węgrzyn and Klamka 2000) or molecular and DNA-based computing systems (see for example Paun et al. 1998, Węgrzyn 2010) are the most promising directions. Another important trend, the one towards parallel computing (see Czech 2010), also requires the advanced hardware architectures and technologies which will support the development of the beyond state-of-the-art AI methods. For discussion on the status of current trends in AI – refer also to chapter 7 of this monograph – the chapter, whose less formal structure, allows for more speculative description.

Returning to that part of AI, which is formally described in chapter 2, note that in section 2.4, all the methods previously presented in sections 2.2.1, 2.2.2 and 2.3 are illustrated in one practical application, hybrid pattern recognition system. This system uses computer-generated hologram playing the role of the feature extractor, which has been optimized by evolutionary approach with an objective function defined by rough set-based coefficient. The classification of characteristic features obtained form the optimized computer-generated hologram is performed by the artificial neural networks, inspired by the biological nerve systems, as presented in section 2.2.1.

It has been pointed out that neural networks process information in massively parallel way according to connectionist paradigm. The goal of this paradigm is not perfect storing of the training facts and perfect response to these facts, but rather building of the statistical model of the process, which underlies these data. Therefore, the ability to generalize the knowledge known from examples crucial for medical diagnosis (Tadeusiewicz 2009) is characteristic for neural networks. It resembles to some extent (and with clear limitations) the inductive process of perception, cognition and model building performed by human brain.

Out of many neural network architectures and learning algorithms (see Tadeusiewicz 1993, Żurada 1992, Hertz et al. 1991, Korbicz, et al. 1994), there have been described in detail those, which in the opinion of the author have the great impact on the development of the field. These include MLPs and Kohonen's maps as examples of feed-forward networks, and Hopfield's networks as representatives of recurrent networks, which minimize the Lapunov energy function. As learning algorithms, the backpropagation algorithm (classical and with inertial term) applied to MLP networks was presented because of its universality and the role in the rebirth of the ANN area after Minsky and Papert (1969) critique. Some of the most important modifications of the WTM algorithm used in Kohonen's SOMs were also given as examples of unsupervised learning.

Note, that MPL and SOM networks are also the most representative examples of the two main types of classifiers. The first type is the regressive classifier, in which learning is based on minimization of error between required and actual outputs. The author's work with this type of the classifiers include applications in medical diagnosis (Cyran et al. 1997, Ciemniewski et al. 1997), speech recognition (Cyran and Podeszwa 1999), image recognition (Cyran and Jaroszewicz 2000, Cyran 2003), and stylometry (Stańczyk and Cyran 2007a – compare with Stańczyk and Cyran 2007b, where rough set-based approach is applied in the same application). In the context of the learning algorithm, the regressive classifier is represented by MLP trained for example by backpropagation, but also by conjugate gradient method, or by variable metric method. The latter is the implementation of Newton's second order minimization, which uses information about the curvature of the error functional kept in Hessian matrix.

The second type is classification by the choice of the closest neighborhood. In this case, the classification of the unknown input vector is performed by comparison of its similarity to the set of pattern vectors, called prototypes of given classes. Then, the input vector is recognized as belonging to class the most resembling the prototype. Learning of such networks, the example o which is Kohonen's SOM, is based on the appropriate formation of prototypes (Danoeux 1997).

In section 2.2.2 the evolutionary computing was presented, as a very universal method for the optimization, including multi-objective optimization. Many examples of effective applications of genetic algorithms are given by Goldberg (1989), and those, which are solved with evolutionary algorithms with more complex representations, are presented by Michalewicz (1992). It is also worthwhile to mention that many hybrid ANN-evolutionary systems have been developed. One of the most obvious applications of the evolutionary optimization in the field of ANNs is learning of neural networks. Another interesting hybrid, which uses ANNs as a tool for generation of the initial population for further evolutionary computation, is reported by Rutkowska et al. (1999).

In the field of rough set theory, section 2.3.1 presents its major generalizations and modifications, and on that background, section 2.3.2 presents author's modification of the indiscernibility relation, used in the theory of rough sets. This theory has been successfully applied to many machine learning and artificial intelligence oriented problems. However, it is well known limitation of this theory, that it processes continuous attributes in an unnatural way. To support more natural processing, the modification of indiscernibility relation has been proposed (2.3:3), such that the indiscernibility relation remains the equivalence relation, but the processing of continuous attributes becomes more natural.

This modification introduces the information about the structure to classically unstructured collection of the attributes that the relation is dependent on. It has been shown

that the classical relation is the special case of the modified version, therefore proposed modification can be recognized as being more general (yet, not as general, as indiscernibility relations, which are no longer equivalence relations). Remarkably, proposed generalization is equivalently valid for classical theory of rough sets, as well as for the variable precision model, predominantly used in machine learning applied to huge data sets.

Proposed in section 2.3.2 modification of the indiscernibility relation, introduces the flexibility in definition of particular special case, which is most natural to given application. In the case of real-valued attributes, our modification allows for performing multidimensional cluster analysis, contrary to multiple one-dimensional analyses, required by the classical form. In majority of cases, the cluster analysis should be performed in a space, generated by all attributes. This corresponds to a family **C** composed of one set (*card* (**C**) = 1), containing all conditional attributes, and is the opposite case, compared to the classical relation, assuming that family **C** is composed of one-element disjoint sets, and therefore, satisfying equation *card* (**C**) = *card* (*C*). However, other less extreme cases are allowed as well and, in an experimental study presented in section 2.4, there was used a family $\mathbf{C} = \{C_R, C_W\}$, composed of two sets containing 8 elements, each. Such structure seems to be natural for application having two-way architecture, like HRWD-based feature extractor.

In this latter application, the modification allowed to improve the recognition abilities by reducing the normalized decision error by 6.5%, if a system, optimized with classical indiscernibility relation, is treated as the reference. One should notice, that this improvement is achieved with respect to a reference, being already optimized solution, which makes any further improvement difficult. Obtained results experimentally confirm the claims concerning sub optimality of solutions optimized with classical indiscernibility relation.

As it has been already mentioned, the experiment presented in section 2.4 is an illustration of application of proposed AI-based methodology to hybrid pattern recognizer. The combined, connectionist and rule-based, approach in this application is reported in Cyran (2005a) and the comparison of the two approaches in evolutionary optimization of the feature space is given in Cyran (2007c). While the mentioned hybrid pattern recognizer uses in very natural way modification of the indiscernibility relation presented in section 2.3.2, it should be stressed that this modification (see Cyran and Stańczyk 2007a, Cyran 2008b) can find many more applications in rough set-based machine learning, since it gives natural way of processing real-valued attributes, within a rough set based formalism.

Certainly there are also limitations. Because some known in rough set theory notions loose their meaning, when the modified relation is to be applied, therefore, if for any reason, they are supposed to play relevant role in a problem, the proposed modification can be hardly applied in any other than classical special case form. One prominent example concerns so called basic sets in a universe *U*, defined by the indiscernibility relation, computed with

respect to single attributes, as opposed to modified relation predominantly designed to deal with sets of attributes defining a vector space, used for common cluster analysis.

Despite this limitation, this modification is practically useful, especially in the case of information systems with real-valued conditional attributes representing the vector space $\mathcal{R}^N$, such as systems of non syntactic pattern recognition. The experimental example belongs to this class of problems and illustrates the potential of modified indiscernibility relation for processing real-valued data in a rough set based theory.

Concerning other rough set model modifications, DRSA is no doubt a powerful tool for information retrieval from data representing the preference ordered criteria. However, if the problem can be naturally reduced to discrete criteria and binary preference-ordered decision, then this sophisticated construction, designed to be as universal as possible, can be less appropriate than QDRSA, proposed by the author and presented in section 2.3.3 as an efficient approach dedicated for such type of applications.

The real-world illustration, described in chapter 4, section 4.3.3, is an example that such class of applications is of practical value, at least in all problems with automatic interpretation of a battery of statistical tests. The genetic example with neutrality tests is only one of them. Certainly, many other areas exist which have similar properties from the information retrieval point of view. In presented there illustration, the information preserved in the combination of neutrality tests has been retrieved by author's method called QDRSA most efficiently than with CRSA and DRSA (see section 4.3.3 for details).

3. POPULATION GENETICS MODELS

3.1. Foundations

Individuals in a natural population exhibit considerable similarity as well as certain degree of morphological difference. The similarity come from the fact that these individuals share the same genetic architecture, while their difference is caused by a number of factors, from the variation of genes to different environmental impacts. The genetic variation, and in particular the dynamics of this variation is studied by population genetics, the science which formulates general principles (for example Hardy-Weinberg law) and models (for example the Wright-Fisher model or the coalescent model) used by theories of evolution on molecular level.

For a sufficiently large population composed of diploid individuals which mate at random and reproduce in non-overlapping generations the frequency of different but selectively neutral alleles (i.e. alternative forms of a gene which correspond to phenotypes having identical fitness) is constant from generation to generation if the effect of the mutation can be neglected.

Theorem 3.1:1 (Hardy-Weinberg equilibrium law, after Hartl and Clark 1997)

Assuming that the frequency of the first allele A_1 is p and the frequency of the second allele A_2 is $q = 1 - p$, the frequency of the three possible genotypes: the homozygote A_1A_1, the heterozygote A_1A_2, and the homozygote A_2A_2, is in equilibrium after one round of random mating. Moreover, the frequency of the homozygote A_1A_1 is p^2, the frequency of the heterozygote A_1A_2 is $2pq$, and the frequency of the homozygote A_2A_2 is q^2.

Proof

From the assumption about independent choice of both gametes it follows that during one round of random mating the proability of the birth of the homozygote A_1A_1 is equal $p \times p = p^2$, the probability of the birth of the homozygote A_2A_2 is equal $q \times q = q^2$, and the frequency of the birth of the heterozygote A_1A_2 is equal $1 - p^2 - q^2 = (p^2 + 2pq + q^2) - p^2 - q^2 = 2pq$.

From the assumption of random mating in large populations (ideally infinite size population), mentioned probabilities are equal to the frequencies in population, and therefore, the result follows.

■

The Hardy-Weinberg law has an important consequence for the fate of rare alleles. Suppose A_2 allele is rare, that is $q = 1 - p$ is small. The question arises: are A_2 alleles more likely to be in A_2A_2 homozygotes or in A_1A_2 heterozygotes? From Hardy-Weinberg law it follows that he ratio of the latter to the former is

$$\frac{2pq}{q^2} = \frac{2p}{q} \approx \frac{2}{q} \tag{3.1:1}$$

Short analysis of (1) (Gillespie 1998) gives an evidence that the rare alleles much more frequently occupy heterozygotes than homozygotes. It suggests that the fate of rare alleles is tied to their dominance relationship with the common A_1 allele, and that therefore, dominance is an important factor in evolution.

The generalization of the Hardy-Weinberg equilibrium for multiple alleles is also valid. Suppose there are k alleles at the locus. Then, under the same assumptions as before, the genotypic frequencies will reach equilibrium in one generation, and the Hardy-Weinberg proportions can be calculated by the following expansion (Fu 2003):

$$\left(p_1A_1 + p_2A_2 + ... + p_kA_k\right)^2 = \sum_{i=1}^{k} p_i^2 A_iA_i + \sum_{j=2}^{k}\sum_{i=1}^{j-1} 2p_ip_jA_iA_j \tag{3.1:2}$$

The generalized Hardy-Weinberg law is used to define the homozygosity G and heterozygosity H of the locus.

Definition 3.1:1 (Homozygosity, after Gillespie 1998)

The homozygosity G of the locus is defined as

$$G = \sum_{i=1}^{k} p_i^2 . \tag{3.1:3}$$

▬

Definition 3.1:2 (Heterozygosity, after Gillespie 1998)

The heterozygosity of the locus H is defined as

$$H = 1 - G = 1 - \sum_{i=1}^{k} p_i^2 . \tag{3.1:4}$$

▬

Consider now two loci, locus A with two alleles A_1 and A_2, and locus B with two alleles B_1 and B_2. The four mentioned above alleles have frequencies $p_1, p_2,$ and q_1, q_2, respectively. If

these loci are located in different chromosomes then they are unlinked physically (and therefore statistically) because they segregate independently according to the Mendel's law. When they are located on the same chromosome, they are physically linked, but they can be in linkage equilibrium (i.e. statistically independent). Certainly they also can be in the linkage disequilibrium.

Two loci A and B can produce four gametes: A_1B_1, A_1B_2, A_2B_1, and A_2B_2 with frequencies P_{11}, P_{12}, P_{21}, and P_{22}, respectively. These loci are in the linkage equilibrium if $P_{11} = p_1q_1$, $P_{12} = p_1q_2$, $P_{21} = p_2q_1$, and $P_{22} = p_2q_2$. This means that the association between alleles from the two loci are at random, i.e. what happens at one locus is independent of what happens on the other locus. If the above equations do not hold, then the two loci are in the linkage disequilibrium.

The most commonly used measure of the linkage disequilibrium is (Fu 2003)

$$D = P_{11} - p_1q_1 = -(P_{12} - p_1q_2) = -(P_{21} - p_2q_1) = P_{22} - p_2q_2 = P_{11}P_{22} - P_{12}P_{21}. \quad (3.1:5)$$

The dynamics of the linkage disequilibrium is shaped by the recombination. In fact, an individual can produce four types of gametes for two loci with two alleles at each locus due to recombination. Therefore, the recombination rate r between two loci, i.e., the probability that the two chromosomes recombine at a point between the two loci, has got the strong influence on the time course of the linkage disequilibrium.

Let $P_{ij}(t)$ be the value of P_{ij} at generation t and $D(t)$ be the value of D at generation t. A randomly selected allele from individual of generation $t + 1$ is one randomly selected from the gene pool of generation t. The gene pool of generation t is a collection of all the gametes produced by individuals of that generation. Typically, each individual is assumed to contribute an infinite number of gametes according to their probabilities. Thus it is convenient to regard a gene pool as of infinite size regardless of the population size. This is so called Random Union of Gametes (RUG) model, which is equivalent to Random Union of Zygotes (RUZ) model, but because of simplicity the RUG is more preferred.

Therefore, the probability that the genotype A_iB_j can be produced in the next generation is given by (Fu 2003)

$$P_{ij}(t+1) = \sum_k P(A_iB_j|I_k)P(I_k). \quad (3.1:6)$$

where I_k is a diploid individual of the certain type k (there can be $[(n^2 + 1)\, n^2\,] / 2$ such types for two loci with n allele at each locus. After some algebra, the equation (6), having in mind (5), results in the following four formulas for two loci with 2 alleles each

$$P_{11}(t+1) = P_{11}(t) - rD(t), \quad (3.1:7)$$

$$P_{12}(t+1) = P_{12}(t) + rD(t), \quad (3.1:8)$$

$$P_{21}(t+1) = P_{21}(t) + rD(t), \tag{3.1:9}$$

$$P_{22}(t+1) = P_{22}(t) - rD(t). \tag{3.1:10}$$

Note also that

$$\begin{aligned} D(t+1) &= P_{11}(t+1)P_{22}(t+1) - P_{12}(t+1)P_{21}(t+1) = \\ &= [P_{11}(t) - rD(t)][P_{22}(t) - rD(t)][P_{12}(t) + rD(t)][P_{21}(t) + rD(t)] = . \\ &= (1-r)D(t). \end{aligned} \tag{3.1:11}$$

Providing that the population at time $t = 0$ has the linkage disequilibrium $D(0)$, from (11) the geometric decay of the linkage disequilibrium is expected with parameter (1 – r)

$$D(t) = (1-r)^t D(0). \tag{3.1:12}$$

This has the implication that for any two loci with $r > 0$, linkage equilibrium will eventually be reached but the speed to equilibrium, depending on the recombination rate r, can be so slow (the smaller recombination rate, the slower decay) that the linkage disequilibrium can be practically maintained in population between closely located loci. For testing linkage disequilibrium (see Fu 2003) the X^2 statistic having χ^2 distribution with one degree of freedom can be used

$$X^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum \frac{(O_{ij} - nP_{ij})^2}{nP_{ij}} = \sum \frac{(O_{ij} - n\hat{p}_i\hat{q}_j)^2}{n\hat{p}_i\hat{q}_j}, \tag{3.1:13}$$

where n is a sample size, O_{ij} and E_{ij} are the observed and expected numbers of alleles of the type A_iB_j, respectively, and the frequencies p_i and q_i can be estimated by

$$\hat{p}_i = \frac{1}{n}\sum_j O_{ij}\,, \tag{3.1:14}$$

$$\hat{q}_i = \frac{1}{n}\sum_j O_{ji}\,. \tag{3.1:15}$$

For two loci with two alleles at each locus, equation (13) can be transformed to (Fu 2003)

$$X^2 = \frac{nD^2}{\hat{p}_i\hat{p}_j\hat{q}_i\hat{q}_j}\,. \tag{3.1:16}$$

Many methods used in population genetic studies are based on the Wright-Fisher model of genetic drift, which assumes a multinomial sampling scheme from generation to generation and thus a binomial distribution of the number of offspring of any particular chromosome. For large populations, the binomial distribution is approximated by the Poisson distribution. These issues are discussed below in more detail, starting from the simple Wright-Fisher model of the genetic drift without mutation and selection (section 3.2), and then in more complex models with mutation and selection (sections 3.3 and 3.4).

For a sample of DNA sequences not undergoing recombination, it is assumed that all these sequences are descendants of an ancestral chromosome existing some generations ago. This time is referred to as the time to coalescence of the whole sample. Similarly we define the time to coalescence of two chromosomes randomly drawn from the sample. The notion of coalescence is the basis for the coalescent theory described in section 3.5.

3.2. Genetic drift and the Wright-Fisher model

The Hardy-Weinberg's implication concerning the constancy of the allele frequency from generation to generation is based on the assumption of the infinite size of the population. Despite the fact, that in real populations this assumption is never satisfied, the populations with large population sizes often conform to the Hardy-Weinberg equilibrium, what was experimentally confirmed many times using X^2 statistic having approximately χ^2 distribution with one degree of freedom. Nevertheless, there are cases when populations go through the periods, called the bottlenecks, when they have relatively small sizes. During such periods, a phenomenon called the genetic drift has got important influence on shaping the allele frequency.

Random genetic drift is a term to describe changes in allele frequencies due to chance in reproduction in populations of finite size. The consequence of the genetic drift is that in randomly mating diploid populations of finite sizes N, in the absence of mutation and selection, one out of $2N$ gametes will eventually be fixed, and all but one gametes will be lost. The time to achieve that state is called the time to fixation. It can be shown that the mean time (in terms of number of generations) for fixing a gamete is $4N$ generations (Hartl and Clark 1997). Hence it is clear that the speed of fixation depends on the population size. Moreover it is easy to demonstrate that each of the gametes has equal chance to be fixed. The probability of fixation of the particular gamete is therefore $1/2N$. In the two-allele situation the probability of fixing allele A_1 is p.

The process of fixation and loss of alleles due to the genetic drift seems to be in a clear opposition to a stable state predicted by the Hardy-Weinberg law. However, evolutionary forces responsible for these two phenomena operate on completely different time scales: Hardy-Weinberg equilibrium is achieved in one generation time-span, whereas the genetic drift requires on average $4N$ generations for fixing the gamete. Therefore the effects of the genetic drift in may population are below level required for detection and that is the reason why many finite size population are confirmed to be in Hardy-Weinberg equilibrium. While the general influence of the genetic drift on evolution is disputable, there are some regions where it is important, to mention bottlenecks and the evolution of rare alleles (often responsible for genetic diseases) as some well known examples.

The genetic drift is predicted by the Wright-Fisher (W-F) model. The W-F model is a model of reproduction and its assumption is that each individual of a new generation is formed by selecting two gametes randomly form the gamete pool of the previous generation. Therefore the W-F model is formulated in terms of the RUG model. Alternatively the W-F model can be described by saying that each allele at generation $t + 1$ is randomly selected from the alleles at generation t with replacement and the population evolves as a Markov chain. From both these definitions it is clear that the W-F model is about how individuals are formed and does not make assumptions about the population size. This implies that population size can vary over generations. The W-F model with constant population size assumes that population size remains constant over generations. It is true that such model is very often used because of its simplicity, however constancy of the population size is not an intrinsic feature of the W-F model.

Suppose there are j copies of allele A_1 and $(N - j)$ copies of allele A_2 at the present generation. This is the two-allele situation with $p = j/2N$ and $q = 1 – j/2N$. Then, the number i of allele A_1 in the next generation can take a value between 0 and $2N$, and it follows the binomial distribution (Fu 2003)

$$P(i|j)=\frac{(2N)!}{i!(2N-i)!}\left(\frac{j}{2N}\right)^{i}\left(1-\frac{j}{2N}\right)^{2N-i}. \tag{3.2:1}$$

From the properties of binomial distribution

$$E\left(\frac{i}{2N}\right)=p \tag{3.2:2}$$

and

$$Var\left(\frac{i}{2N}\right)=\frac{p(1-p)}{2N}. \tag{3.2:3}$$

In multiple allele situation with k different alleles (see Fu 2003), let n_i (t) be the number of allele i at generation t. Then the allele numbers n_i (t+1) follow the multinomial distribution

$$P(n_1(t+1),...,n_k(t+1)|n_1(t),...n_k(t))=$$
$$=\frac{2N!}{n_1(t+1)!...n_k(t+1)!}\left(\frac{n_1(t)}{2N}\right)^{n_1(t+1)}...\left(\frac{n_k(t)}{2N}\right)^{n_k(t+1)} \tag{3.2:4}$$

Treating each chromosome as a different allele, let us denote i_j as the contribution to the next generation by chromosome j. Then (4) can be simplified to

$$P(i_1,...,i_{2N})=\frac{(2N)!}{\prod_j i_j!}\left(\frac{1}{2N}\right)^{2N}. \tag{3.2:5}$$

From binomial distribution properties, it follows that the number of progeny i_j of any particular chromosome j, referred to as the contribution to the next generation from this chromosome, has the properties

$$E(i_j) = 2N\frac{1}{2N} = 1 \tag{3.2:6}$$

and

$$Var(i_j) = 2N\frac{1}{2N}\left(1-\frac{1}{2N}\right) = 1-\frac{1}{2N}. \tag{3.2:7}$$

Equations (6) and (7) have important implication: each particular chromosome is expected to propagate to the next generation with exactly one copy. However, for small populations, this expectation can deviate seriously from the actual number of chromosomes. This discrepancy leads to changes in the allele frequencies and eventually to the extinction of some alleles caused by random genetic drift. Moreover, from (2) it is clear that as population size approaches infinity, the frequency of A_1 approaches p. Similarly, as population size approaches infinity, the genotypic frequencies approach the Hardy-Weinberg proportions.

Consider a reproduction scheme in which the contribution i of a chromosome to the next generation follows the Poisson distribution with mean equal to λ. It follows that

$$P(i) = \frac{e^{-\lambda}\lambda^{i}}{i!}. \tag{3.2:8}$$

Providing that the contribution of different chromosomes is independent of each other and N' is the size of the population at the next generation, then the joint contributions have probability distribution

$$P(i_1,...,i_{2N}) = \frac{e^{-2N'\lambda}\lambda^{2N'}}{\prod_{k=1}^{2N} i_k!}. \tag{3.2:9}$$

If one fixes the population size in the next generation to be N', then the Poisson model is equivalent to the W-F model, as the Poisson distribution conditional on a sum becomes the multinomial distribution. It follows that (Fu 2003)

$$P\left(i_1,...,i_{2N}\middle|\sum_{k=1}^{2N} i_k = 2N'\right) = \frac{P\left(i_1,...,i_{2N},\sum_{k=1}^{2N} i_k = 2N'\right)}{P\left(\sum_{k=1}^{2N} i_k = 2N'\right)} = \frac{(2N')!}{\prod_{k=1}^{2N} i_k!}\left(\frac{1}{2N}\right)^{2N'}. \tag{3.2:10}$$

The characteristic feature of the W-F model without mutation is the decay of heterozygosity due to genetic drift. Before further discussion of this phenomenon, let us give some useful definitions.

Definition 3.2:1 (Alleles identical by origin, after Gillespie 1998)

Alleles differ by origin if they come from the same locus on different chromosomes.

▬

Definition 3.2:2 (Alleles identical by state, after Gillespie 1998)

Alleles are different by state if they have different DNA sequence (when DNA sequences are considered) or different amino-acid sequences (if proteins are considered) or they differ in any particular feature under consideration.

▬

Definition 3.2:3 (Alleles identical by descent, after Gillespie 1998)

Alleles differ by descent if they do not share a common ancestor allele.

▬

Note, that formally two alleles are never different by descent, as they always share a remote common ancestor. However, if this ancestor is more than, say 10 generations in the past, for practical reasons, we consider two alleles as different by descent. Note also, that two alleles different by descent may or may not be different by state due to mutation.

Definition 3.2:4 (Coefficient G, after Gillespie 1998)

Let us define coefficient G as a probability that two alleles different by origin (i.e. drawn at random from the population without replacement) are identical by state.

▬

Definition 3.2:5 (Coefficient H, after Gillespie 1998)

Let us define coefficient H as a probability that two alleles different by origin (i.e. drawn at random from the population without replacement) are different by state.

▬

Lemma 3.2:1 (after Gillespie 1998)

The value of G after one round of random mating, G', as a function of the current value, is

$$G' = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)G. \qquad (3.2:11)$$

Proof

These allele are assumed to be neutral because of their identity by state. The formula above is derived as the sum of probabilities of two mutually exclusive events. The first event is that which occurs when after one turn of random mating two randomly drawn alleles are

descendants of the same allele in previous generation (i.e. they are identical by origin). The probability of this event is $1/2N$. The second event is that after one round of random mating two randomly drawn alleles are descendants of two alleles in previous generation (probability $1 - 1/2N$) and at the same time these two parent alleles are identical by state (probability G by definition). Therefore the joint probability of the second event is $(1 - 1/2N)\,G$, and the result follows.

■

The time course for G is most easily studied by using $H = 1 - G$, the probability that two randomly drawn alleles are different by state. From Lemma 1, it is easy to show that

$$H' = 1 - G' = \left(1 - \frac{1}{2N}\right) H. \tag{3.2:12}$$

and finally

$$\Delta_N H = -\frac{1}{2N} H \tag{3.2:13}$$

where

$$\Delta_N H = H' - H. \tag{3.2:14}$$

From (13) it is evident that the probability that two alleles are different by state decreases at a rate 1/(2N) each generation. For very large populations this decrease is very slow, nevertheless, the eventual result is that all of the variation is driven from the population by genetic drift. This formal result corroborates with the initial statements about fixation of certain allele in the population with reproduction approximated by W-F model.

The full time course for H is given by (Gillespie 1998)

$$H_t = H_0 \left(1 - \frac{1}{2N}\right)^t \tag{3.2:15}$$

where H_t is H in the generation t. Formula (15) says that the decay of H is geometric. For large populations, genetic drift is a weak evolutionary force, as may be shown by the number of generations required to reduce H by one-half. This number is the value of t that satisfies the equation

$$H_t = H_0 / 2. \tag{3.2:16}$$

Therefore

$$\frac{H_0}{2} = H_0 \left(1 - \frac{1}{2N}\right)^t. \tag{3.2:17}$$

After canceling H_0 from both sides, taking the natural logarithm of both sides and solving for *t*, it follows that (Gillespie 1998)

$$t_{1/2} = \frac{-\ln(2)}{\ln\left(1-\frac{1}{2N}\right)} \approx 2N\ln(2). \tag{3.2:18}$$

Note that G is a measure of genetic variation in the population, which is almost the same as homozygosity *G* defined in Definition 3.1:1 by equation (3.1:3). The difference is only in drawing two alleles without (for G) and with (for *G*) replacement. It can be shown that (Gillespie 1998)

$$G = \frac{1}{2N} + \left(1-\frac{1}{2N}\right)G = G'. \tag{3.2:19}$$

When there is no variation then G = 1, when every allele is different by state from every other allele, then G = 0. Analogously to G and *G*, probability H is similar to the heterozygosity of the population *H* defined in Definition 3.1:2 by equation (3.1:4).

$$H = \left(1-\frac{1}{2N}\right)H = H'. \tag{3.2:20}$$

Therefore the process of the decay of H is also the process of the decay of heterozygosity *H*.

The W-F model can be also used to define the effective population size of the population. Whereas often this term denotes the number of breeding individuals in the population, in population genetic it has got special meaning.

Definition 3.2:6 (Effective population size, after Fu 2003)

The effective population size N_e, is the size of an ideal population evolving according to the W-F model that has the dame amount of randomness, i.e., the same magnitude of random genetic drift, as the actual population.

■

Magnitude of the random genetic drift can be related to the probability that two randomly selected alleles come from the same allele at previous generation, or to the variance of the allele frequency, or to the speed of fixation of alleles. Consequently, it is possible to define inbreeding effective population size, variance effective population size, and eigenvalue effective population size (Ewens 2003) being the leading eigenvalue of the matrix of transitions from generation *t* to *t* + 1. Out of these three, the first definition is the most natural.

For a diploid population, the probability P that two randomly chosen alleles come from the same allele in the previous generation satisfies $P = 1/(2N)$. Hence, N is related to P as $N = 1/(2P)$, and the inbreeding effective population size $N_{e\_inbreeding}$ of diploid population is computed as (Fu 2003)

$$N_{e\_inbreeding} = \frac{1}{2P}, \tag{3.2:21}$$

what is a basis for the definition, as given below.

Definition 3.2:7 (Inbreeding effective population size, after Fu 2003)

Inbreeding effective population size $N_{e\_inbreeding}$ of diploid population is the reciprocal of twice the probability that two randomly chosen chromosomes come from the same chromosome in the previous generation.

▬

Even if the inbreeding effective population size is defined for only two generations, it is often convenient to define effective population size over more generations. In fact, it is possible to say about short-term effective population size, defined for a short period of time and closely tracking the dynamics of population size, and long-term effective population size which is a sort of average of effective population sizes over a long period of time. The reason for that latter is great simplification of mathematics, for the W-F model with variable short-term effective population size can be modeled by the W-F model with constant long-term effective population size.

Definition 3.2:8 (Long-term effective population size, after Gillespie 1998)

The long-term effective population size N_e is a size of the idealized W-F population whose rate of decay of heterozygosity is the same as that of the considered population.

▬

As Definition 8 says, the concept of long-term effective population size is based on the decay of heterozygosity H at a rate $1/2N_e$ in an ideal W-F population mimicking the rate of decay in a real population with fluctuating population size $N(i)$ indexed by the generation number, no matter how complex the reproduction scheme. The first step is to take into consideration the real, and possibly complex, reproduction scheme and to estimate $N_{e\_inbreeding}(i)$ based on (21). For simplicity let us denote $N_{e\_inbreeding}(i)$ in what is going by N_i.

Theorem 3.2:1 (Long-term effective population size approximation, after Gillespie 1998)

The long-term effective population size N_e is given by the harmonic mean of the short-term effective population sizes N_i

$$N_e \approx \frac{1}{\frac{1}{t}\sum_{i=0}^{t-1}\frac{1}{N_i}}. \tag{3.2:22}$$

Proof (after Gillespie 1998)

Since, for population with variable sizes N_i, instead of (15) the following holds

$$H_t = H_0 \prod_{i=0}^{t-1}\left(1-\frac{1}{2N_i}\right), \tag{3.2:23}$$

hence the long-term effective population size N_e satisfies the equation

$$H_0\left(1-\frac{1}{2N_e}\right)^t = H_0\prod_{i=0}^{t-1}\left(1-\frac{1}{2N_i}\right). \tag{3.2:24}$$

Solving (24) for N_e by canceling H_0, and approximating the product of terms that are close to one, results in the equation

$$\exp\left(-\frac{t}{2N_e}\right) \approx \exp\left(-\sum_{i=0}^{t-1}\frac{1}{2N_i}\right). \tag{3.2:25}$$

Finally, by equating the exponents in (25) and solving for N_e, the result follows.

■

The consequence of the fact that harmonic mean is influenced more by small values than by larger ones, is that populations which underwent bottlenecks have much reduced heterozygosity as compared to simple expectations based on their actual census size.

3.3. Mutation

Genetic drift is an evolutionary force removing genetic variation from populations. The evolutionary force with opposite effects is caused by mutation. The interaction of these two forces leads to mutation-drift equilibrium as it will be shown below. To start discussing mutation, note that it is caused by not perfect copying the DNA sequences between in the reproduction. Therefore, on molecular level the mutation is caused by single nucleotide change, which is the basis for single nucleotide polymorphism (SNP), insertions, deletions including those characteristic to short tandem repeats called microsatellites, as well as other DNA rearrangements.

All these molecular types of mutations can be approximated in population genetics by one of two models: infinite alleles model (IAM) and infinite sites model (ISM). The first assumes that the new mutation creates a new allele not going into details of the nature of genes composed of sequences of nucleotides. The latter assumes that genes are composed of long

nucleotide sequences and the new mutation changes one of them at place mutation occurred never before.

It may seem that this latter assumption is very realistic and therefore ISM, as more realistically depicting the nature, is better approximation of the real process. However, if instead of locus having infinite length, the typical locus of 1000bp is considered, the model can have difficulties with the restricted number of possible alleles (equal to the number of base pairs in the locus). The IAM, not going into details of the organization of a gene more closely resembles the number of possible alleles (effectively infinite with 1000bp long locus). This is the reason why this model is still relevant in population genetics, despite less accurate describing the structure of genes. Note, that in the derivation of the mutation-drift equilibrium both, IAM and ISM models can be used.

Lemma 3.3:1 (Two alleles identical by state in W-F with mutation model, after Hartl and Clark 1997, Gillespie 1998)

Assume that the population of the size $2N$ is subject for the mutation occurring at a rate μ per locus per generation. Then two alleles randomly drawn (without replacement) from the next generation are identical by state with probability G' given by

$$G' = \left\lfloor \frac{1}{2N} + \left(1 - \frac{1}{2N}\right) G \right\rfloor (1-\mu)^2 . \tag{3.3:1}$$

Proof

Note that equation (1) is a product of probability of drawing two chromosomes which are identical by state in W-F model without mutation as given by equation 3.2:11 in Lemma 3.2:1, and the probability of no mutation occurred to any of them in a model with mutation. Hence, the result follows.

■

Theorem 3.3:1 (Mutation-drift equilibrium heterozygosity $\hat{H}$, after Gillespie 1998)

The mutation-drift equilibrium heterozygosity $\hat{H}$ is given by

$$\hat{H} = \frac{4N\mu}{1+4N\mu} . \tag{3.3:2}$$

Proof

Using reasonable approximation of equation (1) in Lemma 3.2:1 by eliminating from summation terms proportional to μ^2 and $\mu/2N$ (μ is typically 10^{-5} or less and $2N$ is typically 10^4 dependent on conditions considered), the equation (1) can be rearranged to

$$G' \approx G + \frac{1}{2N} - \frac{G}{2N} - 2\mu G. \tag{3.3:3}$$

From (3), after some algebra it follows that

$$\Delta H = -\frac{1}{2N} H + 2\mu (1 - H). \tag{3.3:4}$$

Note that the change of heterozygosity in (4) is a sum of negative change $\Delta_N H$ due to genetic drift only, as defined in (3.2:13), and positive change $\Delta_\mu H$ due to mutation only, given by

$$\Delta_\mu H = 2\mu (1 - H). \tag{3.3:5}$$

In equilibrium, when heterozygosity is not changing, equation (4) results in

$$\frac{1}{2N} H = 2\mu (1 - H), \tag{3.3:6}$$

and after some algebra, the result follows.

■

Theorem 3.3:2 (Mutation-drift equilibrium homozygosity $\hat{G}$, after Hartl and Clark 1997)

The mutation-drift equilibrium homozygosity $\hat{G}$ is given by

$$\hat{G} = \frac{1}{1 + 4N\mu}. \tag{3.3:7}$$

Proof

From (3) it follows also that

$$\Delta G \approx \frac{1}{2N}(1 - G) - 2\mu G. \tag{3.3:8}$$

In equilibrium, when homozygosity is not changing, equation (8) results in

$$\frac{1}{2N}(1 - G) = 2\mu G, \tag{3.3:9}$$

and after some algebra, the result follows.

■

Definition 3.3:1 (Composite parameter θ, after Ewens 2003)

The product $4N\mu$, which has particular relevance in population genetics is referred to as a composite parameter θ.

▬

Note, that Equations (2) and (7) are dependent only on the composite parameter θ. Therefore the estimates of this parameter can be obtained from Theorem 1 and 2. The graphs of $\hat{G}$ and $\hat{H}$ as functions of $\theta = 4N\mu$ are presented in Figure 1.

Theorem 3.3:3 (Heterozygosity-based estimate of θ)

The heterozygosity based estimate of θ is given by

$$\hat{\theta} = \frac{\hat{H}}{1-\hat{H}}. \tag{3.3:10}$$

Proof

Using Definition 1, the result follows directly from Theorem 1, equation (2).

■

Theorem 3.3:4 (Homozygosity-based estimate of θ)

The homozygosity based estimate of θ is given by

$$\hat{\theta} = \frac{1-\hat{G}}{\hat{G}}. \tag{3.3:11}$$

Proof

Using Definition 1, the result follows directly from Theorem 2, equation (7).

■

The estimate of θ can be also computed from

$$\hat{\theta} = \frac{\hat{H}}{\hat{G}}. \tag{3.3:12}$$

By comparing equations (4) and (8) it is clear that the mutation has similar effect on homozygosity as genetic drift on heterozygosity. In particular, the change of homozygosity is a sum of negative change $\Delta_\mu G$ due to mutation only and positive change $\Delta_N G$ due to genetic drift only, where

$$\Delta_\mu G = -2\mu G \tag{3.3:13}$$

and

$$\Delta_N G = \frac{1}{2N}(1-G). \tag{3.3:14}$$

From (13) it follows that

$$G_t = G_0(1-2\mu)^t \tag{3.3:15}$$

and the value of *t* that satisfies the equation

$$G_t = \frac{1}{2}G_0 \tag{3.3:16}$$

can be computed as $t_{1/2}$ from

$$\frac{1}{2}G_0 = G_0(1-2\mu)^{t_{1/2}}. \tag{3.3:17}$$

Solving (17) with $t_{1/2}$ it follows that

$$\ln\left(\frac{1}{2}\right) = \mathrm{t}_{1/2}\ln(1-2\mu) \tag{3.3:18}$$

and finally

$$\mathrm{t}_{1/2} = \frac{-\ln(2)}{\ln(1-2\mu)} \approx \frac{1}{2\mu}\ln(2). \tag{3.3:19}$$

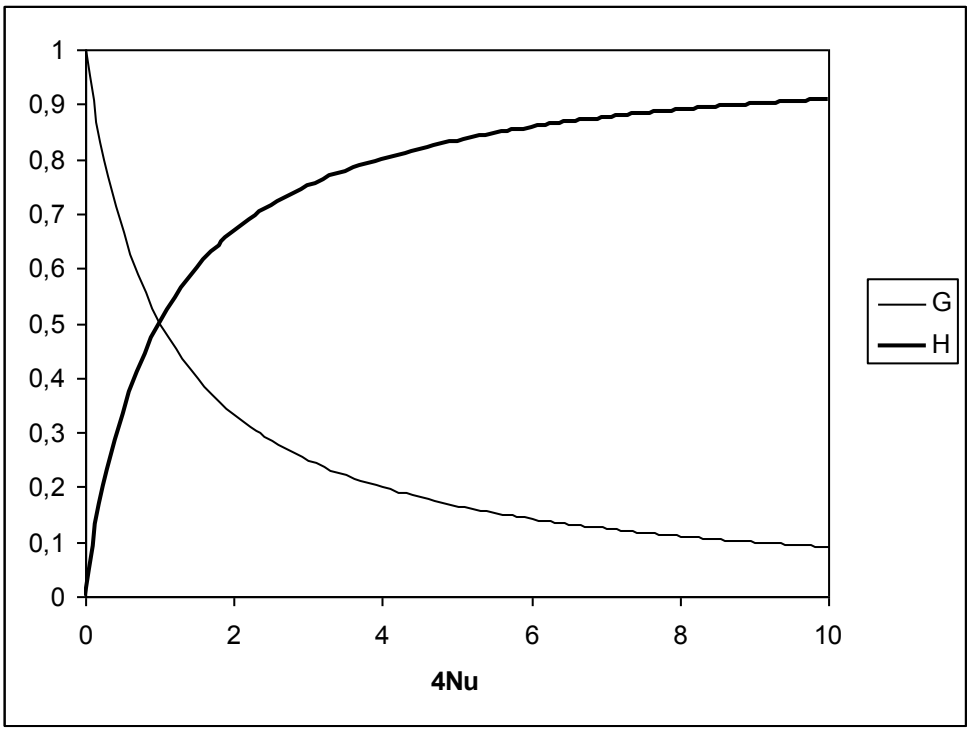


Fig. 3.3:1. Graphs of heterozygosity and homozygosity as functions of composite parameter θ (after Cyran 2008b)

Rys. 3.3:1. Wykresy heterozygotyczności i homozygotyczności w funkcji parametru θ (na podstawie Cyran 2008b)

3.4. Selection

Darwinian evolution would not proceed without natural selection. Therefore, after presenting genetic drift and mutation in previous sections, it is time to consider how the selection operates at the molecular level. The selection model for more than one locus is very complex, and that is the reason why this phenomenon is most often considered for diploid organisms in a one-locus, two allele model.

Definition 3.4:1 (Viability, after Hartl and Clark 1997)

Viability for diploid organisms is the probability that a zygote survives from fertilization to the reproduction.

Suppose, the frequency of the allele A_1 be p, frequency of A_2 be $q = 1 - p$, and viabilities for individuals having genotypes A_1A_1, A_1A_2, and A_2A_2, be w_{11}, w_{12}, and w_{22}, respectively. Consequently, for a population in the Hardy-Weinberg equilibrium, the frequencies of these genotypes at the time of reproduction are $p^2w_{11} / \bar{w}$, $2pqw_{12} / \bar{w}$, and $q^2w_{22} / \bar{w}$, where $\bar{w} = p^2w_{11} + 2pqw_{12} + q^2w_{22}$ is a proportionality constant, denoting the mean viability, and causing the frequencies to add up to 1. Hence, the new allele frequency, p' of the allele A_1, after selection is

$$p' = \frac{p^2 w_{11} + pqw_{12}}{\overline{w}}, \tag{3.4:1}$$

and the change in the allele frequency $\Delta_s p$ is given by

$$\Delta_s p = p' - p = \frac{p^2 w_{11} + pqw_{12} - p\overline{w}}{\overline{w}}, \tag{3.4:2}$$

which, after some algebra, can be rewritten as (Hartl and Clark 1997)

$$\Delta_s p = \frac{pq\left(p\left(w_{11} - w_{12}\right) + q\left(w_{12} - w_{22}\right)\right)}{p^2 w_{11} + 2pqw_{12} + q^2 w_{22}}. \tag{3.4:3}$$

Without loss of generality, suppose the homozygote with two A_1 alleles has larger viability than homozygote with two A_2 allele. The generality is not lost due to arbitrary labeling the alleles. Then, divide numerator and denominator of (3) by w_{11} to obtain the equation expressed in relative viabilities (Gillespie, 1998)

$$\Delta_s p = \frac{pq\left(p\left(1 - \frac{w_{12}}{w_{11}}\right) + q\left(\frac{w_{12}}{w_{11}} - \frac{w_{22}}{w_{11}}\right)\right)}{p^2 + 2pq\frac{w_{12}}{w_{11}} + q^2\frac{w_{22}}{w_{11}}}. \tag{3.4:4}$$

The comparison of (3) and (4) reveals that the dynamics of p do not depend on absolute value of w_{11} but rather on values w_{12} / w_{11} and w_{22} / w_{11}, i.e. values of w_{12} and w_{22} expressed relative to w_{11}. Therefore, it is possible to consider w_{11}, w_{12}, and w_{22} as fitnesses having any values above 0, instead of treating them strictly as viabilities, i.e. probabilities of survival to the reproduction, and having values from 0 to 1. Whatever the range of change of w_{11}, w_{12}, and w_{22}, their values relative to w_{11} are the same, and (4) proves that these relative values really matter for dynamics of allele frequency.

Definition 3.4:2 (Selection coefficient *s*, after Hartl and Clark 1997)

The selection coefficient *s*, given by

$$s = 1 - \frac{w_{22}}{w_{11}} = \frac{w_{11} - w_{22}}{w_{11}}. \tag{3.4:5}$$

is a measure of the difference between the fitnesses of homozygotes relative to the fitness of the homozygote which is more fit.

■

Definition 3.4:3 (Heterozygous effect *s*, after Gillespie 1998)

The heterozygous effect *h*, given by

$$h = \frac{w_{11} - w_{12}}{s w_{11}} = \frac{w_{11} - w_{12}}{w_{11} - w_{22}}. \tag{3.4:6}$$

is the ratio of the differences between fitnesses of heterozygotes and homozygotes.

■

From equation (5) in Definition 2 it follows that $w_{22} / w_{11} = 1 - s$. Similarly, from equation (6) in Definition 3 it follows that $w_{12} / w_{11} = 1 - hs$. Therefore, the relative fitnesses of genotypes A_1A_1, A_1A_2, and A_2A_2 are 1, $1 - hs$, and $1 - s$, respectively. Since as A_1 has been chosen an allele whose homozygote is more fit than A_2 homozygote, the value of selection coefficient *s* is between 0 and 1. The value of heterozygous effect can be arbitrary, however the dominance relation is defined based on this value.

If $h = 0$, then A_1 is dominant, and A_2 is recessive. If $h = 1$, then A_1 is recessive, and A_2 is dominant. Both these extreme case are referred to as complete dominance. If $0 < h < 1$, then there is incomplete dominance, with a special case of $h = 0.5$ in the additive model of selection when the heterozygote has the average fitness of the homozygotes. Finally, if $h < 0$, then there is an overdominance with heterozygote more fit than any homozygote, and if $h > 1$, then there is an underdominance with heterozygote less fit than any homozygote.

Using the relations $w_{22} / w_{11} = 1 - s$, and $w_{12} / w_{11} = 1 - hs$, the equation (4) can be rewritten, as (Gillespie 1998)

$$\Delta_s p = \frac{pqs\big(ph + q(1-h)\big)}{\overline{w}}, \tag{3.4:7}$$

where

$$\overline{w} = 1 - 2pqhs - q^2 s. \tag{3.4:8}$$

Equation (7) can be used to determine the dynamics of the frequency *p*, of the allele A_1.

Definition 3.4:4 (Directional selection, after Gillespie 1998)

Directional selection occurs when fitness of A_1A_1 exceeds that of A_1A_2, which in turn, exceeds that of A_2A_2.

■

Lemma 3.4:1 (Directional selection, after Gillespie 1998)

In the case of incomplete dominance ($0 < h < 1$) the directional selection occurs. With this type of selection the frequency of A_1, p will eventually become one.

Proof

For the incomplete dominance ($0 < h < 1$), from (7) it follows that the sign of the $\Delta_s p$ is always positive, implicating a constant increase of the A_1 allele frequency. Hence, eventually, the frequency of the A_1 allele will approach one, and the result follows.

■

The graph of $\Delta_s p$ as a function of current frequency p is given in Figures 1, 2, and 3, for $s = 0.1$, and different values of heterozygous effect h, taken from the range corresponding to the incomplete dominance.

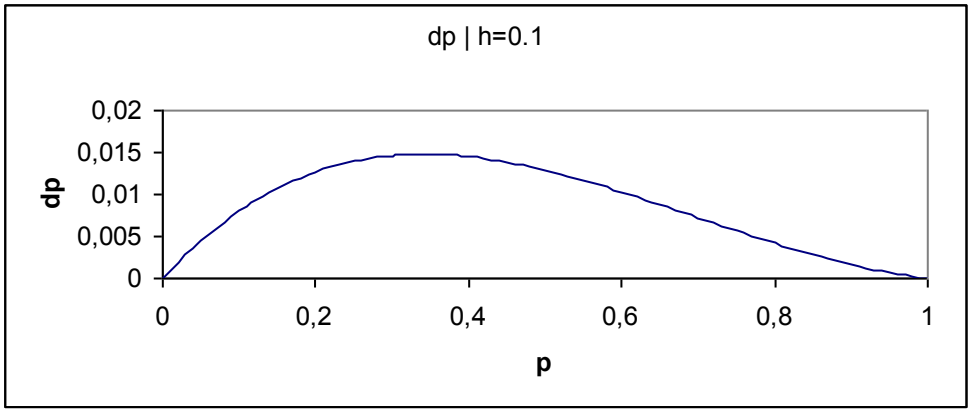


Fig. 3.4:1. Graph of $\Delta_s p$ as a function of p, for directional selection with A_1 almost dominant ($h = 0.1$)
Rys. 3.4:1. Wykres $\Delta_s p$ jako funkcji p, dla selekcji kierunkowej z prawie dominującym A_1 ($h = 0.1$)

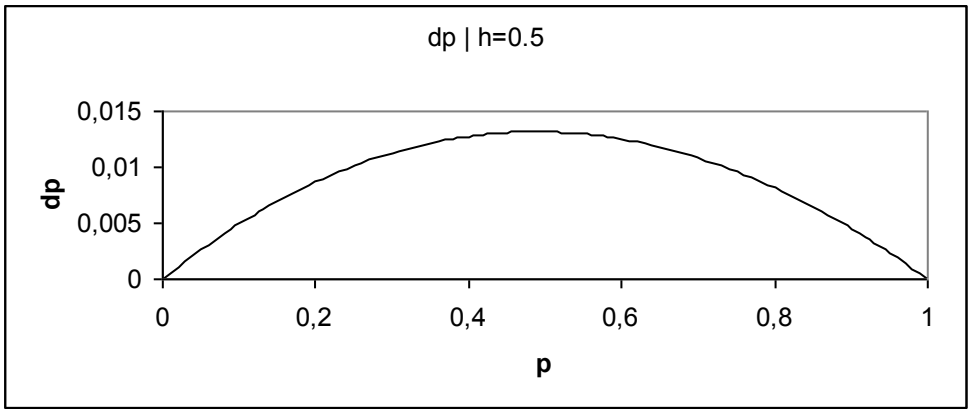


Fig. 3.4:2. Graph of $\Delta_s p$ as a function of p, for additive directional selection model ($h = 0.5$)
Rys. 3.4:2. Wykres $\Delta_s p$ jako funkcji p, dla modelu addytywnej selekcji kierunkowej ($h = 0.5$)

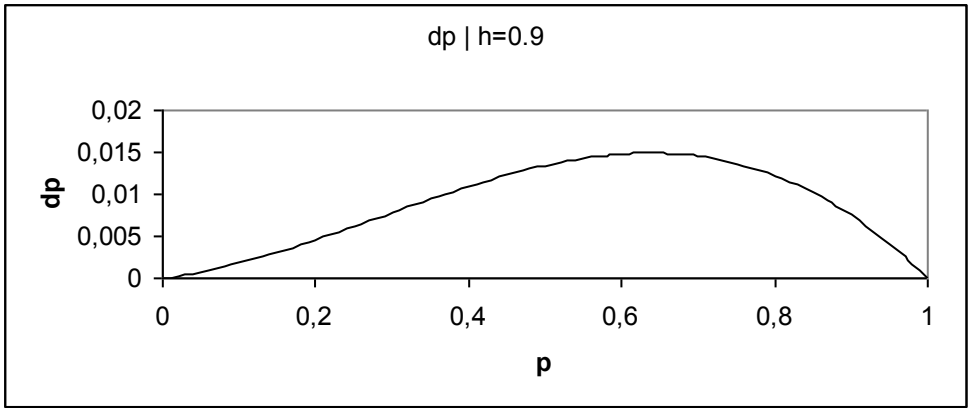


Fig. 3.4:3. Graph of $\Delta_s p$ as a function of p, for directional selection with A_1 almost recessive ($h = 0.9$)
Rys. 3.4:3. Wykres $\Delta_s p$ jako funkcji p, dla selekcji kierunkowej z prawie recesywnym A_1 ($h = 0.9$)

The time course of the frequency p in the directional additive model of selection is presented in Figure 4. Observe the fastest rate of increase when $p(t)$ is around 0.5, i.e., in generations 40 – 60. This observation is in accordance with Figure 2, showing that the highest values of $\Delta_s p$ in this model are for $p = 0.5$.

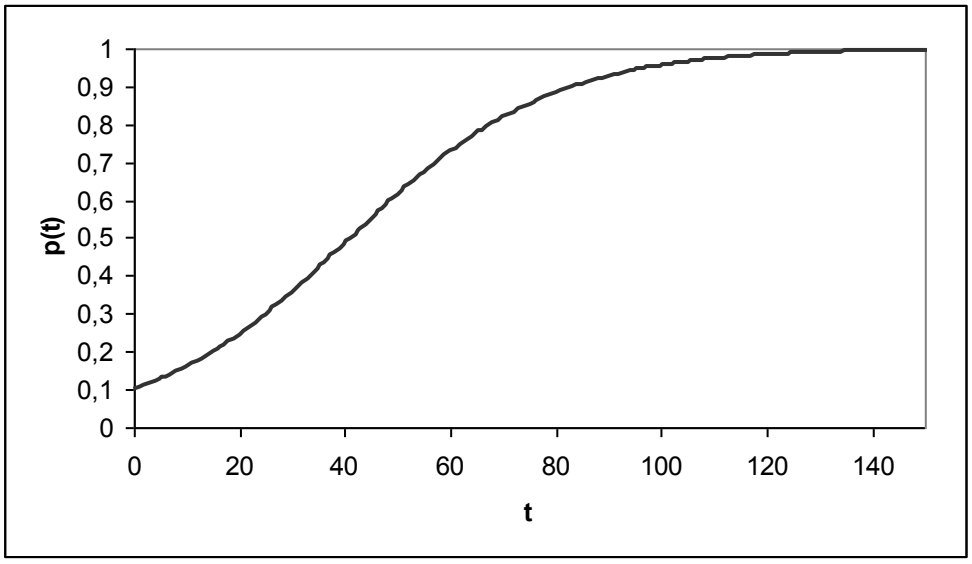


Fig. 3.4:4. Time course of $p(t)$ in the additive, directional selection model ($t(0) = 0.1$, s = 0.1, $h = 0.5$)
Rys. 3.4:4. Wykres $p(t)$ w modelu addytywnej selekcji kierunkowej ($t(0) = 0.1$, s = 0.1, $h = 0.5$)

Definition 3.4:5 (Overdominance selection, after Gillespie 1998)

The second type of selection is called overdominance selection, and it occurs when the heterozygote is more fit that any of the homozygotes.

Lemma 3.4:2 (Overdominance selection, after Gillespie 1998)

The condition required for the overdominance is the negative value of the heterozygous effect. Then the allele frequency achieves stable equilibrium

$$\hat{p} = \frac{h-1}{2h-1}. \tag{3.4:9}$$

Proof

The graphs of $\Delta_s p$ as functions of the current frequency p are given in Figures 5 and 6, for $s = 0.1$, and different values of heterozygous effect h, taken from the range corresponding to overdominance. Since this type of selection is reflected in positive values of $\Delta_s p$ for small frequencies p, and negative values of $\Delta_s p$ for large frequencies p (see Fig. 5 and 6), the ultimate fate of the allele A_1 is the stable equilibrium with frequency such that $\Delta_s p = 0$. From (7) it follows that in equilibrium

$$\hat{p}h + (1-\hat{p})(1-h) = 0, \tag{3.4:10}$$

and thus the result follows.

■

The comparison of Figures 5 and 6 qualitatively reveals that the larger is the absolute value of the heterozygous effect h in the overdominance selection model, the smaller is the frequency at the equilibrium. The equation (9) in Lemma 2 shows that the limit frequency is 0.5, what happens when $h \to -\infty$. Because the above mentioned frequency is being kept in a balanced (or stable) equilibrium, the overdominance selection is often referred to as balancing selection.

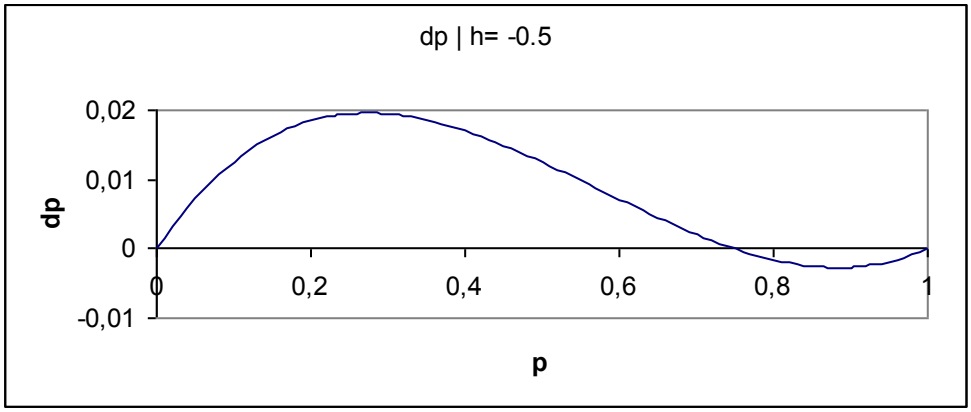


Fig. 3.4:5. Graph of $\Delta_s p$ as a function of p, for balancing selection with $h = -0.5$
Rys. 3.4:5. Wykres $\Delta_s p$ jako funkcji p, dla selekcji balansującej z $h = -0.5$

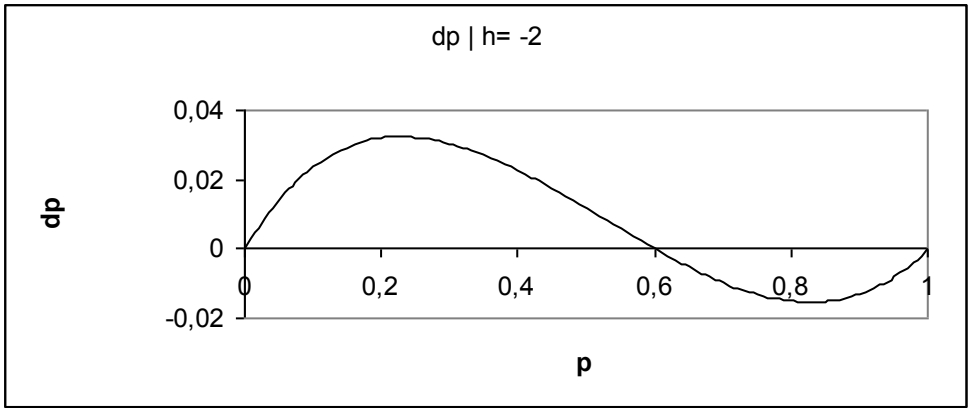


Fig. 3.4:6. Graph of $\Delta_s p$ as a function of p, for balancing selection with $h = -2$
Rys. 3.4:6. Wykres $\Delta_s p$ jako funkcji p, dla selekcji balansującej z $h = -2$

The time course of the frequency p in the overdominance model of selection is presented in Figure 7. Observe that the trajectories of frequencies ultimately settle in the equilibrium no matter from which initial value the evolution starts. This observation is in accordance with Figures 5 and 6, showing the positive values of $\Delta_s p$ for small frequencies and negative values of $\Delta_s p$ for large frequencies of the A_1 allele.

Note, that in the case of balancing selection caused by overdominance mechanism, the mutant allele is kept in the population for a very long time, and sometimes it is even reflected in phenomenon called between-species polymorphism. It is also responsible for keeping in a population strongly deleterious recessive alleles, which in the extreme case can be even lethal when present in homozygotes. The most famous example of this is the allele responsible for sickle cell anemia, abundant in the endemic regions of malaria. Its high frequency in these regions is kept by balancing selection, since the heterozygotes having this allele are partially resistant to malaria (Cavalli-Sforza and Bodmer 1971). The effect is due to the sickling phenomena , making red blood cells less suitable for the *Plasmodium falciparum*, one of four major species responsible for malaria. Similar mechanism is the most probable explanation of the polymorphism shared by humans and chimpanzees in the ATM gene as explained in section 4.3.2 after discovery by the author and co-workers of the balancing selection operating at this gene (Cyran, Polańska, and Kimmel 2004).

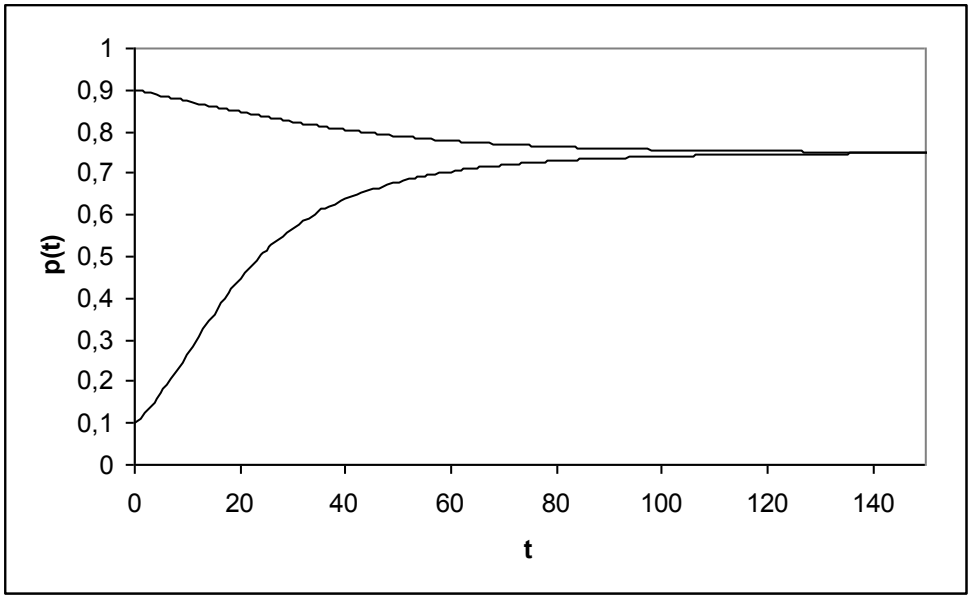


Fig. 3.4:7. Time course of $p(t)$ in the overdominance selection model (s = 0.1, h = – 0.5, $p(0)$ = 0.1 for the bottom curve, $p(0)$ = 0.9 for the upper curve)
Rys. 3.4:7. Wykres $p(t)$ w modelu selekcji ponaddominującej (s = 0.1, h = – 0.5, $p(0)$ = 0.1 dla dolnej krzywej, $p(0)$ = 0.9 dla górnej krzywej)

Definition 3.4:6 (Underdominance selection, after Gillespie 1998)

The third type of selection is called the underdominance selection, as it denotes the situation when the heterozygote is less fit than any of the homozygotes.

■

Lemma 3.4:3 (Underdominance selection, after Gillespie 1998)

The underdominance selection is reflected in the value of the heterozygous effect h greater than one. The ultimate fate of the allele A_1 is dependent on its initial frequency. The underdominance selection model predicts an unstable equilibrium, and the mutant allele is relatively quickly eliminated from the population.

Proof

Note, that his type of selection is reflected in negative values of $\Delta_s p$ for small frequencies p, and positive values of $\Delta_s p$ for large frequencies p (see Fig. 8 and 9). Therefore, the ultimate fate of the allele A_1 is dependent on its initial frequency. If this frequency is below the unstable equilibrium then the allele will be removed from the population. If the initial frequency is above the unstable equilibrium level, then the allele will be fixed. Keeping the equilibrium frequency, although theoretically possible, is unlikely, due to random changes caused by for example random genetic drift. Any deviation from the equilibrium results in the one of the above mentioned scenarios: loss or fixation. Since the mutant allele has typically very small frequency, it is removed from the population, what ends the proof.

■

This type of selection is very rarely, if ever, met in natural populations, however, as it is argued by Gillespie (1998) the chromosomes that differ by translocations and inversions between some closely related species, can be the examples of the underdominance selection taking place before speciation and, to some extent, being responsible for the speciation. The change of the allele frequency under this type of selection can be observed in Figures 8 and 9, which present the graphs of $\Delta_s p$ as functions of the current frequency p, for $s = 0.1$ and different values of heterozygous effect h, taken from the range corresponding to underdominance.

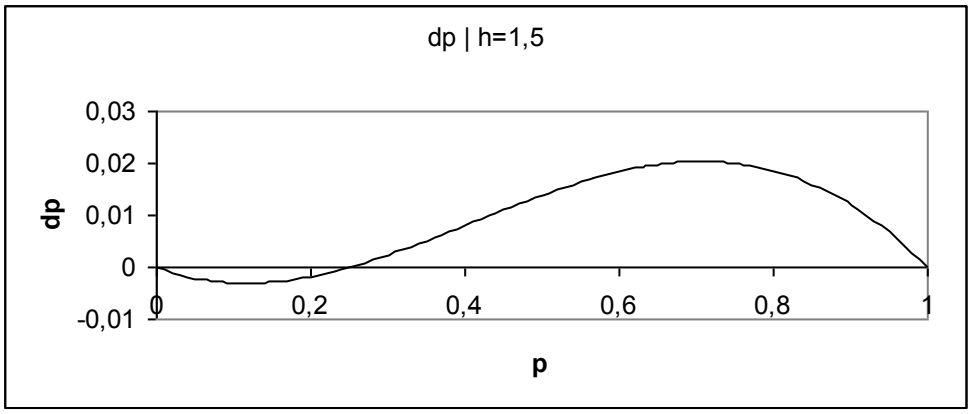


Fig. 3.4:8. Graph of $\Delta_s p$ as a function of p, for underdominance selection with $h = 1.5$
Rys. 3.4:8. Wykres $\Delta_s p$ jako funkcji p, dla selekcji subdominującej z $h = 1.5$

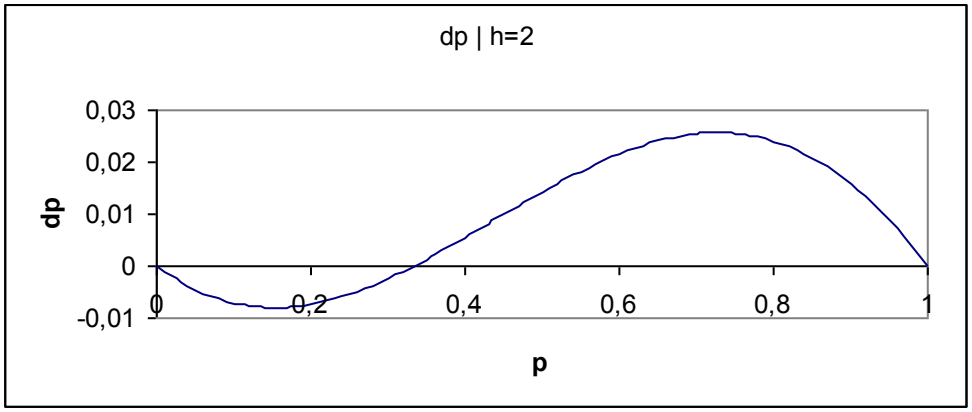


Fig. 3.4:9. Graph of $\Delta_s p$ as a function of p, for underdominance selection with $h = 2$
Rys. 3.4:9. Wykres $\Delta_s p$ jako funkcji p, dla selekcji subdominującej z $h = 2$

The evolution of two alleles whose frequencies have been deviated in opposite directions from the equilibrium is presented in Figure 10.

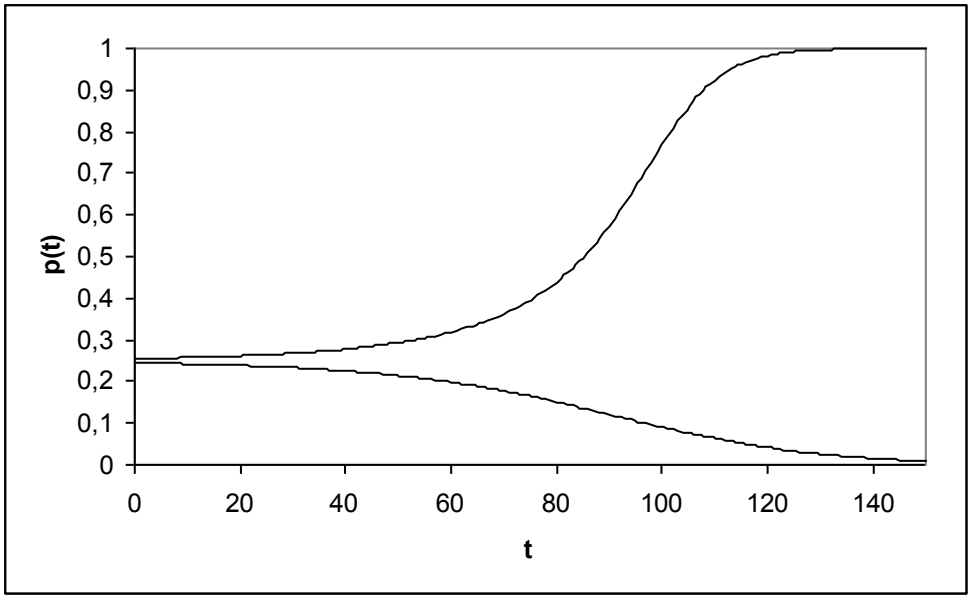


Fig. 3.4:10. Time course of $p(t)$ in the underdominance selection model (s = 0.1, h = 1.5, $p(0) = 0.25 + dp$, for the upper curve, and $p(0) = 0.25 - dp$, for the bottom curve)
Rys. 3.4:10. Wykres $p(t)$ w modelu selekcji ponaddominującej (s = 0.1, h = 1.5, $p(0) = 0.25 + dp$, dla górnej krzywej, oraz $p(0) = 0.25 - dp$, dla dolnej krzywej)

Theorem 3.4:1 (Fate of alleles under selection, after Gillespie 1998)

The final fate of the allele A_1 is determined by heterozygous effect h, while the speed of approaching the limit frequency is determined by the selection coefficient s.

Proof

From Lemma 1, Lemma 2, and Lemma 3 it follows that the final fate of the allele is determined by the type of the selection, which, in turn, is dependent on the value of heterozygous effect h. From equation (7) it is clear that the change of allele frequency is proportional to selection coefficient s, hence the speed of the allele change is dependent on s, what ends the proof.

■

The three types of selection can be observed also by the analysis of the Figure 11. It is a graph of the equilibrium frequency $P_e = \hat{p}$ given by Lemma 2, equation (9), as a function of the heterozygous effect h. The actual frequencies $\hat{p}$ can have only values between 0 and 1. Therefore, the values of h between 0 and 1 (directional selection), corresponding to P_e being less than 0 or more than 1, correspond to no actual equilibrium frequency $\hat{p}$. Moreover, the graph presented in Figure 11 reveals that in the case of overdominance selection ($h < 0$), the equilibrium frequency is always greater than 0.5. Similarly, in the case of underdominance selection ($h > 1$), the equilibrium frequency is always less than 0.5.

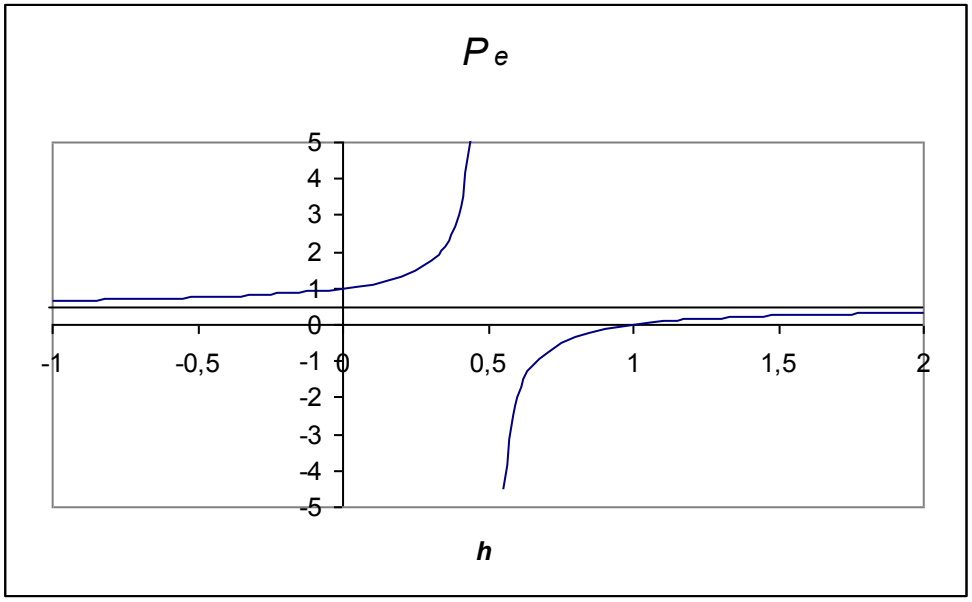


Fig. 3.4:11. The graph of the equilibrium frequency P_e as a function of the heterozygous effect h
Rys. 3.4:11. Wykres częstości równowagi P_e jako funkcji efektu heterozygotycznego h

Theorem 3.4:2 (Mean Fitness Increase Theorem, after Ewens 2003, Gillespie 1998)

The mean fitness in any population will always increase as a result of the natural selection, and the frequency change $\Delta_s p$ is always in that direction, which increases the mean fitness $\bar{w}$. Moreover, the rate of the frequency change in p is proportional to pq , tha laatter reflecting the genetic variation in population. Therefore, the evolution, which is pushed forward by natural selection, operates with the highest rate in populations having high variation.

Proof

While the three different selection types have radically different dynamics (see Lemma 1, Lemma 2, Lemma 3, and Theorem 1), there exists a fundamental law discovered by Wright and Fisher relating the change of the allele frequency $\Delta_s p$ in any type of selection with the slope of the mean fitness viewed as a function of the allele frequency p. The law is expressed by the equation (Gillespie 1998)

$$\Delta_s p = \frac{pq}{2\bar{w}} \frac{d\bar{w}}{dp} \tag{3.4:11}$$

which shows that under natural selection the change of the allele frequency $\Delta_s p$ is proportional to the slope of the mean fitness function. The graphs of the mean fitness $\bar{w}$ as a function of the frequency p can be made based on equation (8). They are presented in Figure

12 for three types of selection and the value of selection coefficient $s = 0.1$. Since the factor $pq / 2\bar{w}$ in (11) is always positive, it is enough to consider what happen for two possible signs of the slope of the mean fitness. (A) If the slope of the mean fitness is positive, it corresponds to positive $\Delta_s p$. Hence, the frequency p increases and it results in the increase of the mean fitness, as its derivative with respect to p is also positive. (B) If the slope of the mean fitness is negative, so is $\Delta_s p$. Therefore, the frequency p will decrease, resulting in the increase of the mean fitness, as its derivative with respect to p is negative. Based on partial results for (A) and (B) situations, the theorem holds.

■

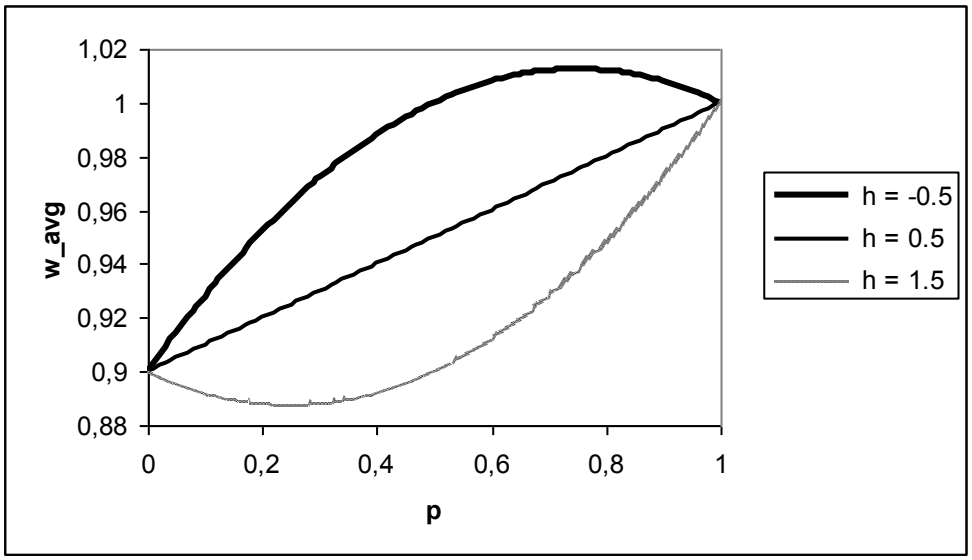


Fig. 3.4:12. The graph of the mean fitness $\bar{w}$, as a function of p for three kinds of selection
Rys. 3.4:12. Wykres średniego dopasowania $\bar{w}$ jako funkcja p dla trzech rodzajów selekcji

The last issue considered in this section is the problem of mutation-selection balance. Assume, that one-way mutation with intensity μ. occurs from allele A_1 with frequency p to allele A_2 with frequency $q = 1 – p$. Suppose also, that the directional selection takes place on genotypes, and that the new mutations are strongly deleterious. Such mutations are partially recessive ($h < 0.5$), and therefore their effect is often covered by the fact that they are most often encountered in heterozygote. This is due to the fact that large effect deleterious mutations are kept by directional selection in very low frequencies, and alleles with low frequencies occupy the heterozygotes rather than the homozygotes. The ratio of the former to the latter is inversely proportional to the allele frequency, what (3.1:1) clearly demonstrates as a consequence of the Hardy-Weinberg equilibrium.

In the absence of selection, the frequency of the allele A_2 in the next generation, q' satisfies

$$q' = q + (1-q)\mu \tag{3.4:12}$$

and therefore, taking in mind that q is very small

$$\Delta_\mu q = \mu(1-q) = \mu - \mu q \approx \mu . \tag{3.4:13}$$

Using the fact that the increase of q is equal to the decrease of p, it follows that

$$\Delta_\mu p \approx -\mu . \tag{3.4:14}$$

For very small q, the equation (7), derived for selection working in isolation of mutation, can be approximated by

$$\Delta_s p = \frac{pqs(ph + q(1-h))}{1 - 2pqsh - q^2 s} \approx qhs , \tag{3.4:15}$$

In the mutation-selection equilibrium the effective change of the allele frequency $\Delta p = \Delta_s p + \Delta_\mu p$ must be equal to zero. Therefore, using (14) and (15) it follows that

$$\hat{q} \approx \frac{\mu}{hs} , \tag{3.4:16}$$

The existence of the deleterious alleles in equilibrium decreases the mean fitness of population. This effect is measured by the genetic load L defined as (Gillespie 1998)

$$L = \frac{w_{\max} - \overline{w}}{w_{\max}} , \tag{3.4:17}$$

where w_{max} is the maximum fitness of the genotype observed in the population. The closer mean fitness to the maximum fitness, the less is the genetic load. Based on (8) and on (16), the mean fitness in mutation-selection equilibrium is given by

$$\overline{w} = 1 - 2\hat{p}\hat{q}hs - \hat{q}^2 s \approx 1 - 2\frac{\mu}{hs}hs = 1 - 2\mu , \tag{3.4:18}$$

and the genetic load becomes

$$L = \frac{1 - (1 - 2\mu)}{1} = 2\mu , \tag{3.4:19}$$

In the case of balancing selection the maximum fitness w_{max} is that of heterozygotes and it is equal $1 - hs$. Therefore, assuming no mutation and the equilibrium frequency for balancing selection given by Lemma 2, equation (9), the genetic load is equal to (Gillespie 1998, after correction of the algebraic error present in the book)

$$L = \frac{1 - hs - (1 - 2\hat{p}\hat{q}hs - \hat{q}^2 s)}{1 - hs} = \frac{\hat{q}^2 s - hs(\hat{p}^2 + \hat{q}^2)}{1 - hs} = \frac{hs(1-h)}{(2h-1)(1-hs)} . \tag{3.4:20}$$

The graph of the genetic load as a function of h in an overdominance case for $s = 0.1$ is presented in Figure 13.

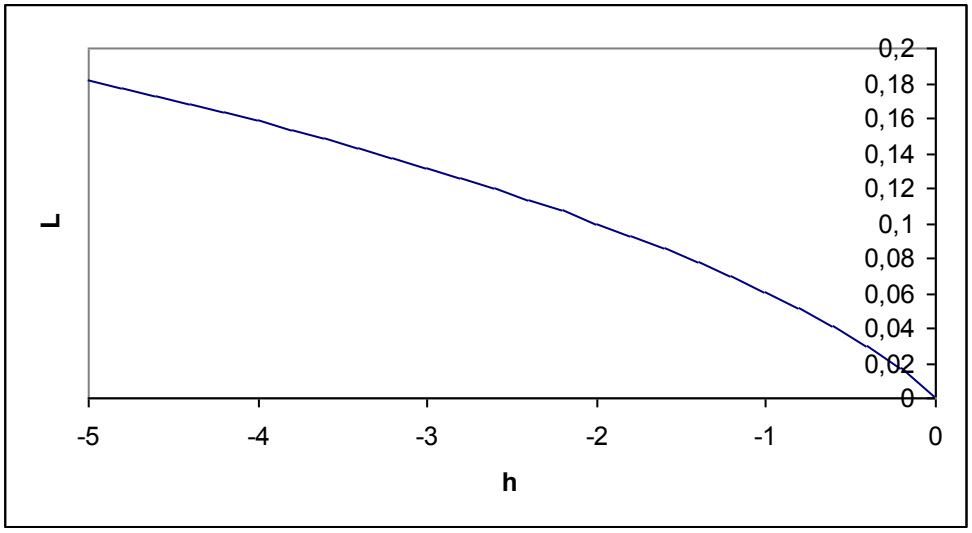


Fig. 3.4:13. The graph of the genetic load as a function of heterozygous effect
Rys. 3.4:13. Wykres ładunku genetycznego jako funkcji efektu heterozygotycznego

This graph presents an counterintuitive result, as the genetic load is greater at the equilibrium frequency $\hat{p}$, when the mean fitness is maximum (see Fig. 12, overdominance case), as compared to situation when the population is composed of only A_1A_1 homozygotes. For that latter situation the genetic load is zero, nevertheless, the mean fitness, equal to 1, is less than the mean fitness at the equilibrium (equal to $1 - hs$, with negative value of h and positive value of s).

3.5. The coalescent model

Consider a sample of DNA sequences from a locus with no recombination. Looking backward in time, a single sequence that is the ancestor of all these sequences will be eventually found. The ancestral relationship creates a phylogeny of these sequences referred also as to gene genealogy or simply genealogy, and defines the notion of a coalescent. Namely, a coalescent is the lineage of alleles in a sample traced backward in time to the allele which is a most recent common ancestor (MRCA) of the whole sample (Fig. 1). When two arbitrary sequences coalesce, i.e. when the number of lineages in the coalescent is reduced by one, it is called a coalescent event.

Definition 3.5:1 (Coalescent time, after Ewens 2003)

The number of generations between successive coalescent events is called the coalescent time. More specifically, the length of the period during which there were n ancestral alleles (sequences) is called n-coalescent time and denoted by T_n. This period is sometimes referred to as the state n of the coalescent process.

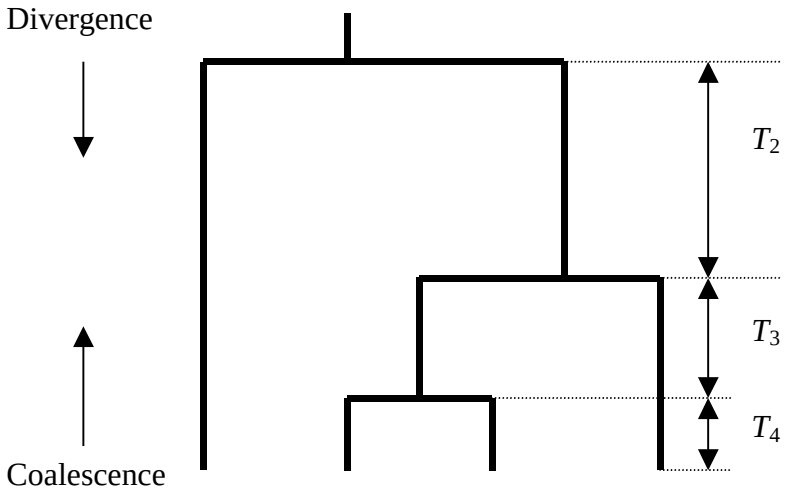


Fig. 3.5:1. The coalescent of four sequences
Rys. 3.5:1. Koalescent dla czterech sekwencji

The Wright-Fisher model implies that at any generation two randomly selected sequences can have the same ancestral sequence in the previous generation. Therefore, when a coalescent event occurs, it is between two randomly selected sequences. By following the process until the MRCA of the whole sample is found, the ancestral relationship among sequences (genealogy) is created which is essentially a random tree. Note, however that not every genealogy has the same probability. In the above tree let us introduce the classification of branches as internal and external. An external branch is the one that connects directly to a sequence in a sample, otherwise the branch is said to be internal.

A random tree of gene genealogy can be also generated by a top-down approach. Starting with the MRCA of the whole sample and splitting it into two descendant lineages creates the first divergence event (see Fig. 1). Then, by random picking one of the lineages and splitting it into two lineages, the second divergence event is modeled. Repetition of this process until there are n lineages leads to the genealogy of n sequences in a sample. Remarkably, this top-down generation of the genealogy leads to the random tree which has the same statistical properties as the one generated by coalescence (Fu, 2003). This top-down generation of genealogy can be applied to compute the probability of a given genealogy. Tajima (1983) proved that the probability P of a genealogy of n sequences with s branching points that lead to exactly two descendant sequences in the sample is given by

$$P = \frac{2^{n-1-s}}{(n-1)!}. \tag{3.5:1}$$

The description of the coalescent time distributions will be started by considering the Wright-Fisher model for a smallest sample exhibiting effects of the genetic drift, i.e. sample composed of two chromosomes. The model assumes a population of haploid individuals (for example mtDNA sequences), which at time $t \geq 0$ has the size N_t. Since multinomial sampling from a given generation's gene pool is assumed, two individuals at generation $t + 1$ are the descendants of the single member of generation t with probability $p_t = 1/N_t$.

Consequently, with probability $q_t = 1- p_t$ they are descendants of two different members. This is reflected in the following distribution of the time to coalescence T_{2c} of two randomly drawn chromosomes in a population with variable size N_t (Bobrowski and Kimmel 2004)

$$P(T_{2c} = t) = \prod_{k=T-t}^{T-1} q_k - \prod_{k=T-t-1}^{T-1} q_k = p_{T-t-1} \prod_{k=T-t}^{T-1} q_k , \tag{3.5:2}$$

where T denotes the number of generations under consideration, and for mathematical consistency $q_{-1} = 0$ and $p_{-1} = 1$.

Apart from its simplicity such model is attractive because it can be easily applied to genetic data. After scaling by the mutation rate μ, the average pairwise mutation difference within a sample corresponds to the expected value of the coalescence time T_{2c} in the model. Moreover, the discrete nature of generations makes it easy to simulate the demography of the model. Therefore, using Monte-Carlo techniques it is possible to estimate unconditional coalescence distribution by averaging conditional on realizations N_t, the distributions given by (2).

Moreover, if we consider a population of constant population size then it is possible to derive algebraically the expected value of the coalescent time. When necessary, the actual fluctuating population size can be approximated by the long-term inbreeding effective population size N_e given by equation (3.2:22) in Theorem 3.2:1. Then the probability p_2 that two randomly selected sequences come from a single ancestral sequence in the previous generation is (Fu 2003)

$$p_2 = \frac{1}{2N_e} . \tag{3.5:3}$$

Hence, the probability q_2 that they come from different ancestral sequences is given by formula

$$q_2 = 1 - p_2 = 1 - \frac{1}{2N_e} . \tag{3.5:4}$$

Given that these sequences come from different parent sequences at generation $t – 1$, the probability that they still come from different ancestral sequences at generation t is also equal to q_2, and the probability that they coalesce is p_2. Therefore, it follows that the probability

$P(T_2 = t)$ that the two sequences come from a single ancestral sequence $T_2 = t$ generations ago is (Fu 2003)

$$P(T_2 = t) = q_2^{t-1} p_2 = \left(1 - \frac{1}{2N_e}\right)^{t-1} \frac{1}{2N_e}. \tag{3.5:5}$$

Theorem 3.5:1 (Coalescent time for two chromosomes, after Fu 2003)

The coalescent time T_2 , i.e. the waiting time until the next coalescent event occurs between two sequences, has the following properties

$$E(T_2) = 2N_e, \quad Var(T_2) \approx (2N)^2. \tag{3.5:6}$$

Proof

Formula (5) specifies the probability distribution of T_2. It follows that the coalescent time T_2 has got geometric distribution with probability of success $p_2 = 1/2N_e$. Since the mean of the geometric distribution is the reciprocal of the probability of success, the first part of the equation (6) follows. From the properties of the geometric distribution, and taking into account that $2N_e >> 1$, it follows that the variance of the coalescent time T_2 is approximately given by

$$Var(T_2) = \frac{q_2}{p_2^2} = \frac{1 - \frac{1}{2N_e}}{\frac{1}{(2N_e)^2}} = 2N_e(2N_e - 1) \approx (2N_e)^2 \tag{3.5:7}$$

what ends the proof.

■

Note that $e^{-x} \approx 1 - x$, when x is small. Since $1/2N_e$ is quite small in natural populations, the distribution of T_2 given by (5) can be approximated by an exponential distribution with probability density function $f(T_2)$ which satisfies (Fu 2003)

$$f(T_2) = \left(e^{-\frac{1}{2N_e}}\right)^t \frac{1}{2N_e} = \frac{1}{2N_e} e^{-\frac{1}{2N_e}t}. \tag{3.5:8}$$

From the properties of the exponential distribution with parameter $\lambda = 1/2N_e$, it follows that the expected value $E(T_2) = 1/\lambda = 2N_e$, and the variance $Var(T_2) = \lambda^{-2} = (2N_e)^2$. In continuous approximation given by (8) the coalescent time is often scaled so that one unit corresponds to $4N_e$ generations. Consider $T_2' = T_2/(4N_e)$. Then the distribution of T_2' is

$$f(T_2') = 2e^{-2t}. \tag{3.5:9}$$

The distribution of the time to coalescence can be computed also for a sample composed of more than two chromosomes. Consider the genealogy of a sample of n sequences taken

from a population of diploid individuals. From (4) it follows that probability that two particular sequences from a sample do not coalesce is $(2N_e - 1) / 2N_e$. It shows that there are a total of $2N_e$ possible ancestors for the second sequence, but only $(2N_e - 1)$ that are different from the ancestor of the first sequence. Similarly, the probability that the third sequence does not coalesce with none of the two sequences, given that these two sequences have different ancestors, is $(2N_e - 2) / 2N_e$. Therefore the total probability that the first three sequences do not share an ancestor is $(2N_e - 1) / 2N_e \times (2N_e - 2) / 2N_e$. This reasoning can be generalized, and the probability q_n that there is no coalescence in one generation for the n sequences is (Gillespie 1998)

$$\begin{aligned} q_n &= \frac{2N_e - 1}{2N_e} \times \frac{2N_e - 2}{2N_e} \times \ldots \times \frac{2N_e - (n-1)}{2N_e} = \\ &= \left(1 - \frac{1}{2N_e}\right)\left(1 - \frac{2}{2N_e}\right)\ldots\left(1 - \frac{n-1}{2N_e}\right) \end{aligned} . \tag{3.5:10}$$

Expanding the product results in (Fu 2003)

$$\begin{aligned} q_n &= 1 - \frac{1}{2N_e} - \frac{2}{2N_e} - \ldots - \frac{n-1}{2N_e} + O\left(\frac{1}{(2N_e)^2}\right) \approx \\ &\approx 1 - \frac{n(n-1)}{4N_e}. \end{aligned} . \tag{3.5:11}$$

Therefore, the probability $p_n = 1 - q_n$ that there is a coalescence among n sequences is

$$p_n \approx \frac{n(n-1)}{4N_e}. \tag{3.5:12}$$

Since there are $n(n-1)/2$ pairs of sequences in a sample of n sequences, it is clear that the approximation given by (12) assumes that no multiple coalescence occurs in one generation. This approximation is valid when $n(n-1) << 4N_e$.

Having probability of the coalescence p_n in one generation it is possible to compute the distribution of the waiting time for the coalescent event, when the coalescent is in n state, i.e., the distribution of the n-coalescent time T_n. Note that the probability that $T_n = t$ is given by

$$P(T_n = t) = \frac{n(n-1)}{4N_e}\left(1 - \frac{n(n-1)}{4N_e}\right)^{t-1}. \tag{3.5:13}$$

Theorem 3.5:2 (Coalescent time for n chromosomes, after Fu 2003)

The expected value and the variance of the n-coalescent time satisfy

$$E(T_n) = \frac{4N_e}{n(n-1)}, \quad Var(T_n) \approx \left(\frac{4N_e}{n(n-1)}\right)^2. \tag{3.5:14}$$

Proof

Equation (13) indicates that *n*-coalescent time T_n has got geometric distribution with the probability of success $p_t = n(n-1) / 4N_e$. Therefore, its expected value is the reciprocal of the probability of the success, and the first part of equation (14) holds. Moreover, for the number of chromosomes satisfying $n(n-1) << 4N_e$, the following approximation is true

$$Var(T_n) = \frac{q_n}{p_n^2} = \frac{4N_e}{n(n-1)} \times \frac{4N_e}{n(n-1)} \times \left(1 - \frac{n(n-1)}{4N_e}\right) \approx \approx \left(\frac{4N_e}{n(n-1)}\right)^2 \tag{3.5:15}$$

what ends the proof.

▪

A continuous approximation of the *n*-coalescent time results in probability density function $f(T_n)$ given by (Fu 2003)

$$f(T_n) = \frac{n(n-1)}{4N_e} e^{-\frac{n(n-1)}{4N_e}t}. \tag{3.5:16}$$

The distribution (16) after rescaling T_n so that one unit corresponds to $4N_e$ generations leads to the distribution of $T_n' = T_n / 4N_e$

$$f(T_n') = n(n-1)e^{-n(n-1)t} \tag{3.5:17}$$

with

$$E(T_n') = \frac{1}{n(n-1)} \tag{3.5:18}$$

and

$$Var(T_n') = \frac{1}{n^2(n-1)^2}. \tag{3.5:19}$$

Definition 3.5:2 (Total time in the coalescent, after Fu 2003)

Let us define a total time in the coalescent, T_c, as a random variable which depends on the times T_n through

$$T_c = \sum_{i=2}^{n} iT_i. \tag{3.5:20}$$

▬

Theorem 3.5:3 (Expectation of the total time in the coalescent, after Fu 2003)

The expected value of the total time in the coalescent satisfies

$$E(T_c) = 4N_e a_n, \tag{3.5:21}$$

where

$$a_n = \sum_{i=2}^{n} \frac{1}{i-1}. \tag{3.5:22}$$

Proof

Assuming a_n given by (22), using the first part of the equation (14) in Theorem 2 and the fact that the expectation of the sum of random variables is the sum of the expectations of those variables, it follows that the expected value of T_c satisfies

$$E(T_c) = \sum_{i=2}^{n} iE(T_i) = 4N_e \sum_{i=2}^{n} \frac{1}{i-1} = 4N_e a_n, \tag{3.5:23}$$

what ends the proof.

■

Theorem 3.5:4 (Variance of the total time in the coalescent, after Fu 2003)

The variance of the total time in the coalescent satisfies

$$Var(T_c) = (4N_e)^2 b_n, \tag{3.5:24}$$

where

$$b_n = \sum_{i=2}^{n} \frac{1}{(i-1)^2}. \tag{3.5:25}$$

Proof

We can compute the variance $Var(T_c)$ as

$$Var(T_c) = \sum_{i=2}^{n} i^2 Var(T_i) = (4N_e)^2 \sum_{i=2}^{n} \frac{1}{(i-1)^2} = (4N_e)^2 b_n, \tag{3.5:26}$$

what ends the proof.

■

Note, that in continuous coalescent model, after rescaling, the mean and variance of the T_c' are equal to a_n and b_n, respectively.

Coalescents are useful constructs because they can be used to infer some information about the genealogy of a sample based on the mutations which occur on the lineages. Assume that the number of mutations which occur in a sequence in one generation is a Poisson variable with mean equal to μ, the mutation rate per sequence per generation. Then the probability that there are k mutations in a branch of length l generations is

$$P(k \mid l) = \frac{e^{-l\mu}(l\mu)^k}{k!}. \tag{3.5:27}$$

with moments

$$E(k \mid l) = l\mu \tag{3.5:28}$$

and

$$Var(k \mid l) = l\mu . \tag{3.5:29}$$

In the ISM model, each mutation results in a segregating site in the sample. Moreover, based on (28), for a neutral mutation rate of μ, the expected number of mutations in a coalescent is μT_c.

Theorem 3.5:5 (Estimate of θ using the number of segregating sites, after Gillespie 1998)

The number of the segregating sites in a sample contains enough information to estimate the composite parameter $\theta = 4N_e\mu$, according to

$$\hat{\theta} = \frac{S_n}{a_n} . \tag{3.5:30}$$

Proof

Using (28) in the ISM model, it follows that

$$E(S_n \mid T_c) = \mu T_c . \tag{3.5:31}$$

Therefore, the expected number of segregating sites in a sample $E(S_n)$ is given by

$$\begin{aligned} E(S_n) &= \sum_{T_c} P(T_c) E_{T_x}(S_n) = \\ &= E_{T_c}\{E(S_n \mid T_c)\} = \\ &= E_{T_c}(\mu T_c) = \\ &= \mu 4N_e a_n = \\ &= a_n \theta . \end{aligned} \tag{3.5:32}$$

Equation (33) can be used to estimate the composite parameter $\theta = 4N_e\mu$, and the result follows.

■

Moreover, using (29) in the ISM models results in

$$Var(S_n \mid T_c) = \mu T_c , \tag{3.5:33}$$

and therefore, the variance of the number of segregating sites $Var(S_n)$ can be derived using additionally the fact that for any random variable X, $\mathrm{Var}(X) = E(X^2) - E^2(X)$. Hence,

$$
\begin{aligned}
Var(S_n) &= E_{T_c}\left\{E\left(S_n{}^2 \mid T_c\right)\right\} - E^2(S_n) = \\
&= E_{T_c}\left\{Var(S_n \mid T_c) + E^2(S_n \mid T_c)\right\} - E^2(S_n) = \\
&= E_{T_c}\left\{\mu T_c + (\mu T_c)^2\right\} - E_{T_c}{}^2(\mu T_c) = \\
&= a_n\theta + \mu^2 E_{T_c}\left(T_c{}^2\right) - \mu^2 E_{T_c}{}^2(T_c) = \\
&= a_n\theta + \mu^2\left\{E_{T_c}\left(T_c{}^2\right) - E_{T_c}{}^2(T_c)\right\} = \\
&= a_n\theta + \mu^2 Var(T_c) = \\
&= a_n\theta + \mu^2(4N_e)^2 b_n = \\
&= a_n\theta + b_n\theta^2
\end{aligned}
\qquad (3.5:34)
$$

When using equation (30) from Theorem 5 it is important to realize that it holds only if the mutations are selectively neutral, i.e. in the neutral model of molecular evolution (see section 5.3). While the replacement mutations (i.e. mutations which change the amino acids in a protein) not necessarily are selectively neutral, it is almost sure that the silent mutations (those which do not change amino acids) are neutral. Therefore, it is an often practice that only silent variation is considered, when calculating the number of segregating sites S_n in a sample of n sequences.

Before the more advanced coalescent models will be presented, let us give an example of the use of coalescent to derive the heterozygosity H (i.e. the probability that two alleles which are different by origin are also different by state). The two alleles will be different by state only if a mutation occurs on a lineages leading to their common ancestor. In tracing back the ancestry of two sequences either a coalescence or mutation will occur first. If the first event is coalescence then the alleles must be identical by state; otherwise they are different by state. In any particular generation, the probability of coalescence, according to (3), is $1/2N_e$, while the probability of mutation on any of the two lineages is $1 - (1 - \mu)^2 \approx 2\mu$. The probability that a mutation will occur first is the relative probability of the mutation, and therefore (Gillespie, 1998)

$$
\hat{H} = \frac{2\mu}{2\mu + \frac{1}{2N_e}} = \frac{4N_e\mu}{1 + 4N_e\mu}, \qquad (3.5:35)
$$

which is the result essentially identical to (3:3:2), but derived much easier with the use of coalescent.

Let us assume a coalescent model with variable in time population size N_τ and continuous time τ measured backwards. Suppose also that $\lambda(\tau) = N_0 / N_\tau$ and that τ_{2c} is the time to coalescence of a pair of alleles measured in N_0 generations. Then the tail of the distribution of τ_{2c} is given by

$$P(\tau_{2c} > \tau) = \exp\left[-\int_0^{\tau} \lambda(u)du \right] \tag{3.5:36}$$

which is the continuous analog of (2). To ensure the coalescence, $\lambda(t)$ must satisfy

$$\int_0^{\infty} \lambda(u)du = \infty \tag{3.5:37}$$

For the stochastic N_τ and therefore $\lambda(\tau)$ there should be expectation over the process on the right side of the equation (36). In the context of problems considered in this book, it is also worth to notice that the continuous coalescence model approximates correctly the discrete coalescent model as long as $1\text{-}1/N_\tau \approx \exp(\text{-}1/N_\tau)$ what certainly is not true in the beginning of branching process (branching processes are discussed in section 3.6), when N_τ (denoted as Z_t when it concerns the population size in branching process – refer to section 3.6 for details) is not large and undergoes stochastic fluctuations. Having this in mind it will be easier to understand the shapes of the experimental distributions of the coalescence presented in the section 5.3.

3.6. Branching processes in population biology

The most concise recursive description of a branching process is given in the following definition.

Definition 3.6:1 (Branching process, after Kimmel and Axelrod 2002)

Branching process is the process in which an ancestral individual produces random number of progeny ($X \geq 0$), and then, each of the progeny independently acts as a new ancestor.

■

The graphical representation of the Definition 1 is given in Fig. 1. Consider doubly infinite family of independent identically distributed random variables $\{X_{i,n}\}$, which denote potential numbers of progeny of i-th individual in generation n. Let Z_n be a number of individuals in generation n. Then the number of individuals in generation $n + 1$ is obtained by summation (Kimmel and Axelrod 2002)

$$\begin{aligned} Z_0 &= 1 \\ Z_{n+1} &= \begin{cases} X_{1,n} + \ldots + X_{Z_n,n}; & Z_n > 0, \\ 0; & Z_n = 0, \end{cases} \quad n \geq 1, \end{aligned} \tag{3.6:1}$$

or

$$Z_{n+1} = \sum_{i=1}^{Z_n} X_{in} \,. \tag{3.6:2}$$

The above formula is called a forward equation, and it is the easiest way to explain the Galton-Watson branching processes. Galton-Watson process evolves in discrete time measured by non-negative integer numbers. The numbers of individuals in subsequent generations of Galton-Watson branching process form a time-discrete Markov chain. The forward equation (2) leads to a recurrent formula for probability generating function (PGF) of the number of individuals in a process.

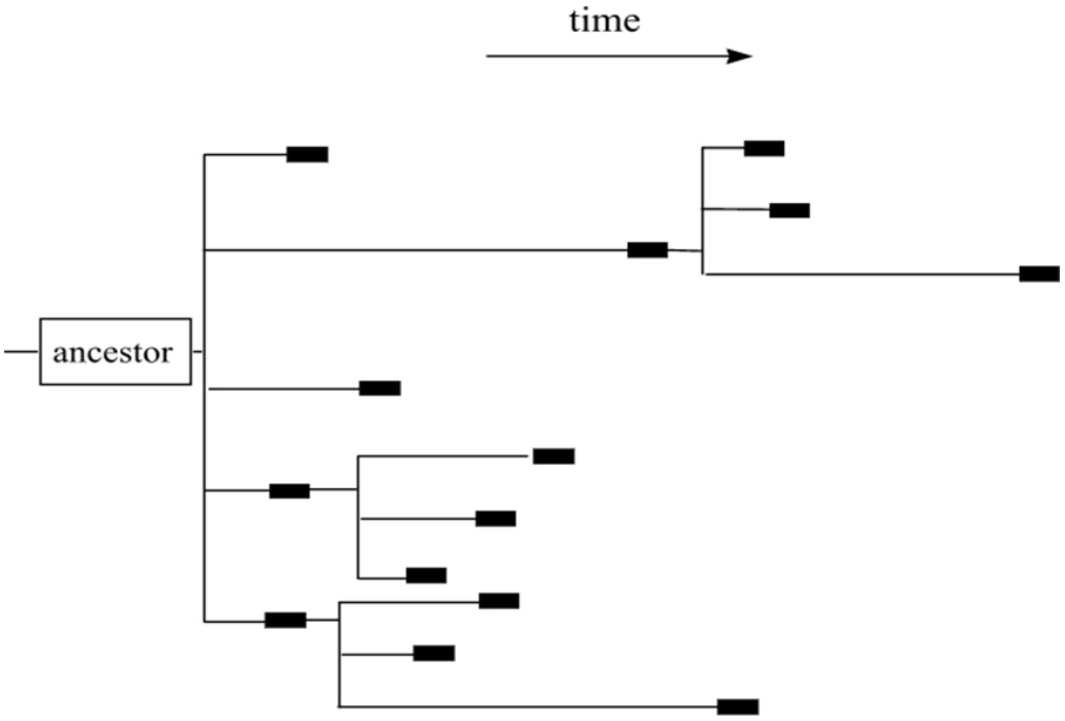


Fig. 3.6:1. The branching process (adapted from Kimmel and Axelrod 2002)
Rys. 3.6:1. Proces gałązkowy (na podstawie Kimmel and Axelrod 2002)

Denote by $f_X(s)$ a PGF for independent identically distributed random variables $\{X_{i,n}\}$, abbreviated to X when no particular individual and/or generation is considered. Then, $f_X(s)$ defined for a symbolic argument $s \in U \equiv [0, 1]$ is given as

$$f_X(s) = E\left(s^X\right) = \sum_{i=1}^{\infty} p_i s^i \tag{3.6:3}$$

The $f_X(s)$ defined by (3) is non-negative and continuous with all derivatives on U, and for non-triviality condition ($p_0 + p_1 < 0$), it is increasing and convex. Moreover (see Feller 1968), the derivatives of $f_X(s)$ satisfy

$$\frac{d^k f_X(0)}{ds^k} = k!\, p_k, \tag{3.6:4}$$

and for proper X, it follows that $f_X(1) = 1$ and the kth factorial moment of X, $\mu_k = E\,(X\,(X-1)\,(X-2)\,\ldots\,(X-k+1)]$ is finite iff $f_X^{(k)}(1-) = \lim_{s\uparrow 1} f_X^{(k)}(s)$ is finite. If this is satisfied, then

$$\mu_k = f_X^{(k)}(1-). \tag{3.6:5}$$

For independent identically distributed non-negative integer random variables $\{X_i, i \geq\}$ and non-negative integer random variable Y, which is independent of sequence $\{X_i\}$, it follows that (see Kimmel and Axelrod 2002) a non-negative integer random variable given by

$$V = \sum_{i=1}^{Y} X_i \tag{3.6:6}$$

has the PGF

$$f_V(s) = f_Y\left(f_{X_i}(s)\right). \tag{3.6:7}$$

Let us denote by $f_n(s)$ the PGF of the number of individuals in the Galton-Watson branching process in generation n. For simplicity let $f_1(s) = f_X(s)$ be denoted by $f(s)$. Then the branching process forward equation (2) and equation (7) lead to the following recurrent formula for $f_{n+1}(s)$

$$f_{n+1}(s) = f_n\left(f_1(s)\right) = f_n\left(f(s)\right). \tag{3.6:8}$$

Since $Z_0 = 1$, which implies $f_0(s) = s$, the above equation yields

$$f_n(s) = \underbrace{f \circ \ldots \circ f}_{n \text{ times}}(s), \tag{3.6:9}$$

which states that the $f_n(s)$ is the nth functional iterate of the progeny PGF $f(s)$.

Similar result can be obtained using the backward approach, which is however more general, and can be applied for arbitrary branching process (not only for Galton-Watson type). The backward approach used a decomposition of the branching process into sub-processes, which are started by the direct offspring (generation 1) of the ancestor (generation 0). In accordance with the branching property these sub-processes are distributed identically as the whole process. This fact is used to derive recurrent relationships for the distribution of the process (Kimmel and Axelrod 2002)

$$Z_{n+1} = \sum_{i=1}^{Z_1} Z_n, \tag{3.6:10}$$

which implies

$$f_{n+1}(s) = f_1\left(f_n(s)\right) = f\left(f_n(s)\right), \tag{3.6:11}$$

and (9) is straightforward.

If we denote by m the mean number of progeny of an individual, then from (5) it follows that

$$m = E(X) = \mu_1 = f'(1-). \tag{3.6:12}$$

Observe also that $m = E(Z_1)$, and

$$E(Z_n) = f_n'(1-). \tag{3.6:13}$$

From the chain rule of differentiation, the derivative of the iterates of a function is a product of derivatives of this function. Hence, the above equation results in

$$E(Z_n) = \underbrace{f'(1-)...f'(1-)}_{n} = m^n. \tag{3.6:14}$$

Application of the chain rule for the second moment yields (Kimmel and Axelrod 2002)

$$Var(Z_n) = \begin{cases} \dfrac{\sigma^2 m^{n-1}(m^n - 1)}{m-1}, & m \neq 1, \\ n\sigma^2, & m = 1 \end{cases} \tag{3.6:15}$$

where $\sigma^2 = \mathrm{Var}(X)$ is the variance of the offspring count. The asymptotic behavior of $f_n(s)$ determines the limit theorems for the process $\{Z_n\}$, as described below.

Theorem 3.6:1 (Extinction probability, after Kimmel and Axelrod 2002)

The extinction probability of the process $\{Z_n\}$ is the smallest non-negative root q of the equation $f(s) = s$ for $s \in [0, 1]$. It is equal to 1 if $m \leq 1$ and it is less than 1 if $m > 1$.

Proof

Since $f(s)$ is a power series with non-negative coefficients $\{p_k\}$, $p_0 + p_1 < 1$, and $f'(1-) = m$, therefore $f(s)$ is strictly convex and increasing in [0,1], $f(0) = p_0$, and $f(1) = 1$. Moreover, if $m \leq 1$ then $f(s) > s$ for $s \in [0,1)$, and if $m > 1$ then $f(s) = s$ has a unique root in $s \in [0,1)$. These properties imply that there exists q being the smallest root of $f(s) = s$ for $s \in [0,1]$. It follows that if $m \leq 1$ then $q = 1$, and if $m > 1$ then $q < 1$.

Additionally, the following statements hold:

a) If $s \in [0, q)$ then $f_n(s) \uparrow q$ as $n \to \infty$,

b) If $s \in (q, 1)$ then $f_n(s) \downarrow q$ as $n \to \infty$,

c) If $s = q$ or $s = 1$ then $f_n(s) = s$ for all n.

Hence, iterates of $f(s)$ converge to q, and as a special case $f_n(0) \uparrow q$ as $n \to \infty$. However, it follows that

$$\begin{aligned} &\lim_{n\to\infty} f_n(0) = \lim_{n\to\infty} P(Z_n = 0) = \lim_{n\to\infty} P(Z_i = 0 \text{ for some } 1 \geq i \geq n) = \\ &P(Z_i = 0 \text{ for some } i \geq 1) = P(\lim_{n\to\infty} Z_n = 0), \end{aligned} \tag{3.6:16}$$

which by definition is the probability that the process becomes extinct for $n \to \infty$, what ends the proof.

■

Consequently, the value of m defines three classes of branching processes, given in the following definitions.

Definition 3.6:2 (supercritical branching process, after Kimmel and Axelrod 2002)

The branching process is called supercritical if the expected number of progeny is greater than one ($m > 1$).

▬

Definition 3.6:3 (critical branching process, after Kimmel and Axelrod 2002)

The branching process is called critical if the expected number of progeny is equal one ($m = 1$).

▬

Definition 3.6:4 (subcritical branching process, after Kimmel and Axelrod 2002)

The branching process is called subcritical if the expected number of progeny is lesser than one ($m < 1$).

▬

The following properties concerning extinction hold for these three cases of branching processes

$$m = E(X)\begin{cases} >1 & (\text{supercritical}) & E[Z_n] = m^n \uparrow \infty & P[\lim_{n\to\infty} Z_n = 0] = q < 1, \\ =1 & (\text{critical}) & E[Z_n] = m^n = 1 & P[\lim_{n\to\infty} Z_n = 0] = 1, \\ <1 & (\text{subcritical}) & E[Z_n] = m^n \downarrow 0 & P[\lim_{n\to\infty} Z_n = 0] = 1. \end{cases} \quad (3.6{:}17)$$

The critical case seems to be a paradox, as the process will ultimately become extinct with probability one, and at the same time the expected number of individuals in a process remains one (does not decrease to zero). The explanation is the growth of the variance. Note, that the linear growth of variance in the critical case given by (15) is consistent with heavy tails of distribution of Z_n for $m = 1$. This growth of variance explains the paradox with critical case, for which $E(Z_n) = 1$ and $\lim_{n\to\infty} P(Z_n = 0) = 1$.

To show this peculiarity visually, Figure 2 provides simulated evolution of a union of 100 critical branching processes (Fig 2a), and 1000 critical branching processes (Fig 2b). Such unions are equivalent to branching process with $Z_0 = 100$ and $Z_0 = 100$, respectively. Observe that branching process in Fig. 2a has become extinct, and that in Fig. 2b will become extinct (for sufficiently large generation number) almost surely, i.e. with probability one, because of the increasing in time variance (this increase is clearly visible in the Fig 2b). Note also that

conditional on non extinction, critical branching processes grow to extremely large sizes (see Fig. 2b) – what at first seems counterintuitive, having in mind that $E(Z_n) = 1$.

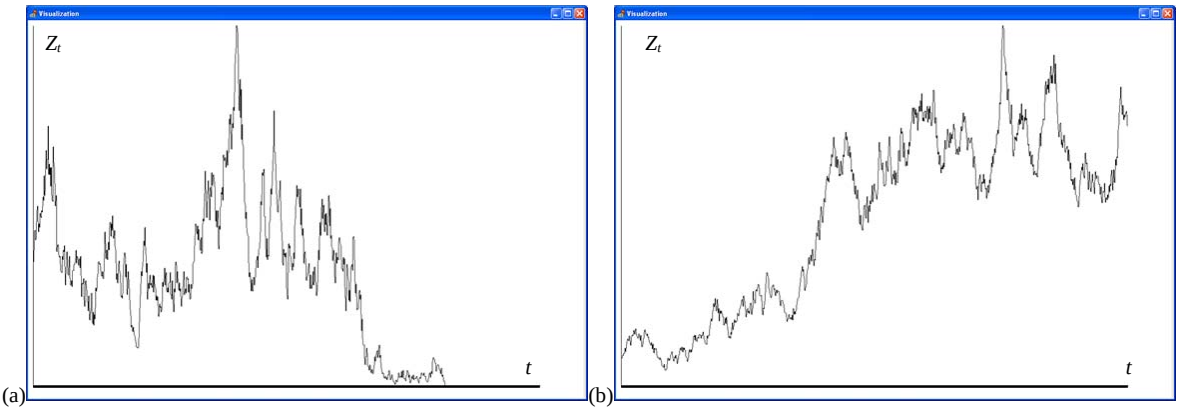


Fig. 3.6:2. Evolution of critical branching process
Rys. 3.6:2. Ewolucja krytycznego procesu gałązkowego

Such instable behavior (extinction or large size, conditional on non extinction) is typical for all classes of branching processes and it generates serious algorithmic challenges in simulating processes for large number of generations, as it is the case in the author's studies described in chapter 5. These studies use branching processes to model evolution of humans of during approximately 10,000 generations (simulations lasting about two weeks on typical PC architecture), what requires writing software with efficient dynamic memory management.

Consider a slightly supercritical time-homogenous (i.e. with parameters not being changed during evolution) branching process $Z_T(t)$ with the expected number of offspring $E(\xi_0) = 1 + \alpha / T + o(1/T)$ and variance $Var(\xi_0) = \sigma^2 + O(1/T)$. For such a model, an asymptotic behavior of the probability $P^x(Z_t > 0)$, where P^x denotes probabilities for the process started by x individuals, is given in following theorems.

Theorem 3.6:2 (Asymptotic behavior of the non-extinction probability, after Cyran and Kimmel 2004b)

If Z_T is a supercritical branching process with $E(\xi_T) = 1+\alpha/T$, $\alpha > 0$, and $\zeta_T = (\xi_T)^2$ is uniformly integrable in T then, as $T \to \infty$

$$1 - P\left[\inf\left\{k > 0 : Z_T(k) = 0\right\} < \infty \mid Z_T(0) = 1\right] \sim \frac{2\alpha}{T\sigma^2}, \tag{3.6:18}$$

where symbol $\sim$ denotes the asymptotic equivalence.

Proof

Denote by q_T the extinction probability P [inf $\{k > 0$: $Z_T(k) = 0\} < \infty \mid Z_T(0) = 1]$. Then by Theorem 1, $q_T = f_T(q_T)$. Taylor expansion of the PGF around 1gives

$$q_T = f_T(1) + (q_T - 1) f_T'(1) + \frac{1}{2}(q_T - 1)^2 [f_T''(1) + R_T(q_T)] =$$
$$= 1 + (q_T - 1) E(\xi_T) + \frac{1}{2}(q_T - 1)^2 f_T''(1) + o\left[(q_T - 1)^2\right] = \quad (3.6:19)$$
$$= 1 + (q_T - 1)\left(1 + \frac{\alpha}{T}\right) + \frac{1}{2}(q_T - 1)^2 \sigma^2 + o\left[(q_T - 1)^2\right]$$

Solving with respect to q_T yields, $q_T = 1 - 2\alpha/(T\sigma^2) + o(q_T - 1)$, which implies the result expressed in equation (18).

■

The left side of equation (18) in Theorem 2 denotes the asymptotic property of the probability of non extinction in infinity of the branching process conditional on the initial size of one individual. More interesting from the evolutionary perspective is the probability of non extinction until given time *t*. By letting *t* to tend to infinity, it is possible to obtain the asymptotic properties of this probability, conditional on the initial size of *x* individuals, as given in Theorem 3, below.

Theorem 3.6:3 (Asymptotic behavior of the non-extinction probability until time *t*, after Cyran and Kimmel 2010)

The probability of non extinction, until any moment t, of the branching process $Z_T(t)$, characterized by the expected number of progeny $E(\xi_T)=1+\alpha/T+o(1/T)$ and the variance $Var(\xi_T)=\sigma^2+O(1/T)$ is given by

$$P^x(Z_t > 0) \sim \frac{2\alpha x}{T\sigma^2\left[1 - \exp\left(-\alpha \frac{t}{T}\right)\right]}, \text{ as } T \to \infty. \quad (3.6:20)$$

Proof

The proof of this theorem based on the convergence in law of the process $\{Z_T(t)/T\}$ to a diffusion can be found in O'Connell (1995).

■

By setting *T* for *t* in (20) we obtain the asymptotic properties of probabilities of non extinction until present

$$P^x(Z_T > 0) \sim \frac{2\alpha x}{T\sigma^2\left[1 - \exp(-\alpha)\right]}, \text{ as } T \to \infty. \quad (3.6:21)$$

Theorem 3.6:4 (Asymptotic property of the expected size of branching process, after Cyran and Kimmel 2004b)

The expected number of individuals (at present time T, as $T\to\infty$) in the slightly supercritical branching process $Z_T(T)$ started by x individuals and extant in T does not depend on x and is proportional to the variance of progeny distribution

$$E\left(Z_T \middle| Z_T > 0, Z_0 = x\right) \sim \frac{\sigma^2 T}{2\alpha}\left[\exp(\alpha) - 1\right], \text{ as } T \to \infty. \tag{3.6:22}$$

Proof

Since

$$E\left(Z_T \mid Z_T > 0, Z_0 = x\right) = \frac{E\left(Z_T \mid Z_0 = x\right)}{P\left(Z_T > 0 \mid Z_0 = x\right)} = \\ = \frac{xE\left(\xi_T\right)^T}{P\left(Z_T > 0 \mid Z_0 = x\right)}. \tag{3.6:23}$$

Therefore using (21) as $T\to\infty$, it follows that

$$E\left(Z_T \mid Z_T > 0, Z_0 = x\right) \sim \frac{x\left(1+\alpha/T\right)^T}{2\alpha\, x} T\sigma^2\left(1-e^{-\alpha}\right) \sim \\ \sim \frac{e^{\alpha} T\sigma^2}{2\alpha}\frac{e^{\alpha}-1}{e^{\alpha}} = \frac{\sigma^2 T\left(e^{\alpha}-1\right)}{2\alpha}, \tag{3.6:24}$$

what ends the proof.

■

Note, that the above result presents, somewhat surprisingly, the asymptotic lack of dependence on Z_0 of the expected value Z_T conditional on non-extinction until present. This effect can be explained, as seen in a proof, by an equal linear influence of $Z_0 = x$ on the unconditional expected value of Z_T and the probability of non extinction until T.

Let us express the time interval [0, T] of a variable t as a unit interval [0, 1] of variable $r = t / T$. Then (O'Connell 1995, corrected in Kimmel and Axelrod 2002), for long times T we have the following equation describing the tail of the distribution of D_T, the time of death of the last common ancestor of the randomly chosen two individuals living at time T, given that we start the population history from x individuals having descendants at T

$$\lim_{T\to\infty} P\left(\frac{D_T}{T} > r \middle| K_0 = x\right) = \frac{2q_r^x}{(x-1)!}\left[(q_r - 1)^{-x}(x-1)! - F(x-1, 1-q_r)\right], \tag{3.6:25}$$

where

$$q_r = \frac{e^{-r\alpha} - e^{-\alpha}}{1 - e^{-\alpha}} \tag{3.6:26}$$

and F: $\mathbf{Z}_+ \times (0,1) \to \mathbf{R}$ is defined as

$$F(n,y)=\frac{\partial^n}{\partial y^n}\left\lfloor \frac{\ln(1-y)}{y^2} \right\rfloor \tag{3.6:27}$$

Moreover, in the O'Connell model it is possible to obtain the following asymptotic formula

$$\begin{aligned} E\left\lfloor \left(1-\frac{D_T}{T}\right) \middle| K_0=1 \right\rfloor &= 1-\int_0^1 P\left(\frac{D_T}{T}>r \middle| K_0=1\right)dr \\ &\xrightarrow{T\to\infty} 1-2\int_0^1 \frac{q_r}{(1-q_r)^2}\left(q_r-1-\ln q_r\right)dr. \end{aligned} \tag{3.6:28}$$

The original O'Connell distribution is continuous but to compare it with the discrete empirical distributions described below, the discretized version is considered, specified by the tail of original distribution computed at points r corresponding to integer values of $t = rT$. For the sake of terminological simplicity, this discretized version of the distribution will be still referred to as to the O'Connell distribution.

3.7. Conclusions

Chapter 3 discussed the population genetics models starting from the Hardy-Weinberg equilibrium given in generalized form in (3.1:2), continuing with the Wright-Fisher model of genetic drift (section 3.2), mutation and selection models (section 3.3, and 3.4, respectively), and ending with a coalescent (section 3.5) and branching processes models (section 3.6).

In context of the second part of the book, dedicated to applications of presented methods in evolution, the implications of particular evolutionary forces on the whole process are of special interest. In that respect, note that the time required for genetic drift to reduce H by one-half is proportional to the population size (3.2:18). For example, in a population having 1million individuals and generation of approximately 20 years the variation is reduced by one-half during 28 million years. Therefore for large populations genetic drift is a very weak evolutionary force. Its interplay with mutation, the evolutionary force with opposite effects, leads to mutation-drift equilibrium, which is discussed in detail in section 3.3 in selectively neutral models.

The predicted values of heterozygosity as homozygosity, as a functions of the composite parameter θ are also considered. By comparison of (3.3:19) with (3.2:18) it is evident that the time of drift-induced reducing heterozygosity by one-half has got the scale N while the time of mutation-induced reducing homozygosity by one-half has got the time-scale μ^{-1}. Since μ^{-1} can be interpreted as the average number of generations required for occurring a single mutation at a locus, therefore the time for homozygosity to be reduced by mutation by one-half can be alternatively expressed in a scale of generations until a typical mutation takes

place. Consequently, the species with shorter generations have got shorter time scale for effects of mutation, and so the evolution in such species proceeds faster than in species with longer generations.

Section 3.4 presents the effects of selection operating at molecular level. Among many results, the one of special importance is that the equilibrium number of deleterious mutations is large enough to have shaped the evolution process towards the evolution of sex, recombination, and the avoidance of interbreeding (Gillespie 1998). Remarkably, the deleterious mutations affect the mean fitness of the population independently of the strength of the selection, decreasing it by an amount 2μ. This somewhat unexpected effect, shown by (3.4:18), can be explained by an equal influence of the selection coefficient s on the deleterious allele frequency as on the mean fitness given that frequency.

Multinomial sampling used in the Wright-Fisher model is also presumed in the coalescent model, a powerful method used for inferring time to the most recent common ancestor (MRCA) in time-backward approach. It has been shown that for large populations the coalescent models are equivalent to diffusion process models whose limiting results depend only on the mean and the variance of offspring number distribution. That issue was discussed in section 3.5 on the basis of coalescent theory. However, the robustness of the coalescent models is valid only for large populations, so for population bottlenecks, like presumably in the case of pre out-of-Africa epoch (see section 5.1), the commonly used diffusion approximation fails, and therefore the BP-based methodologies (described in section 3.6) should be used as it is illustrated in section 5.3 on a real example concerning dating of the Mitochondrial Eve. Problem of criticality of branching processes, addressed in section 3.6, is also basis for the a study concerning the complexity threshold in the early Life (see section 6.2) where the Demetrius-Kimmel (Demetrius et al. 1985, Kimmel and Axelrod 2002) model has been modified by the author (Cyran 2009b) to account for the influence of the phospodiester bond break on dehydrolysis of RNA strands.

Finally, in section 3.6 there was presented a model originally proposed by O'Connell (1995) for dating mitochondrial Eve's death based on a sample of mtDNA of humans and chimpanzees. The implications of the model, as it was shown in section 3.6, and it is further explained in sections 5.3 and 5.4, are far beyond this original application. O'Connell's limiting results are based on the assumption that the population is growing as a slightly supercritical branching process with progeny distributions homogeneous in time. Though these are not quite realistic assumptions, especially time-homogeneity, the model is important as an alternative for the Wright-Fisher model, since it does not assume any particular offspring distribution.

Moreover, asymptotically, for given expected number of offspring, the O'Connell model is independent of the shape of the progeny distribution, and in particular it is independent of its variance as long as this variance is bounded. This property is interesting in the light of classical results where the short-term inbreeding effective population size is proportional to the variance of offspring distribution, and therefore it influences the shape of coalescence distribution. Offspring distribution invariance in O'Connell model is theoretically valid in a limit; however it remained unknown until author's studies described in section 5.3 were performed, how fast, in terms of number of generations, coalescence distributions in real population converge to this asymptotic characteristic. This could have been answered only by time-forward simulation of the full branching process genealogy and then by comparison of actual distributions with limiting results, as presented in section 5.3.

PART II

APPLICATIONS IN EVOLUTIONARY GENETICS

4. THEORY OF NEUTRAL EVOLUTION

4.1. Foundations

Charles Darwin in his famous *On the Origin of Species* (Darwin 1859) tried to explain the variety of forms of the living creatures by the process of evolution pushed forward by natural selection. In section 3.4, focused on the natural selection operating on the molecular level, three kinds of selection have been defined: directional, overdominance, and underdominance. The question arises whether Darwin had thought about all three kinds of selection and if not, which is the selection type considered by him. The analysis of the natural selection as an evolutionary force suggested by Darwin reveals that the type of selection he took in mind was the directional one. However, directional selection can be deleterious, and advantageous. The first type is subject to the selective sweep, the latter is the one, really responsible for the evolution of different species.

Despite the fact that natural selection plays a crucial role in Darwinian evolution, Kimura noticed that there is high cost of evolution pushed by natural selection. That was one of his arguments promoting neutral theory of evolution, and the cause why the neutral evolution was called non-Darwinian evolution (Gillespie 1998). However, even if the neutral theory of molecular evolution claims that most substitutions are due to genetic drift rather than natural selection, Kimura's theory is not in conflict with Darwin's theory. What neutral theory of molecular evolution states is that majority of genetic variation has got no influence on survival of genotypes, however that part of variation which changes the fitness of individuals, is still subject to natural selection with all consequences for evolution process as predicted by Darwin. Discussion of the theory of neutral evolution will be started with the following definition.

Definition 4.1:1 (Average rate of substitution per generation, after Gillespie 1998)

The average rate of substitution per generation, denoted as k is defined as the average number of mutations which will fix in a population each generation.

▬

Theorem 4.1:1 (Average rate of substitution of neutral alleles, after Gillespie 1998)

The rate of substitutions of neutral alleles, k, is equal to the mutation rate to neutral alleles, μ

$$k = \mu . \tag{4.1:1}$$

Proof

Note that the average number of new mutations entering the population each generation is equal $2N\mu$, which is the number of gametes produced each generation times the probability of a mutation in any one of them. Since the probability, that any particular selectively neutral allele will fix in a population due to genetic drift, is equal to the frequency of that allele, it follows that the probability that a particular new mutation will fix is $1/(2N)$. Therefore, of the $2N\mu$ new mutations that enter the population each generation, a fraction, $1/(2N)$, will fix on average. Hence, on average $2N\mu \times 1/(2N) = \mu$ mutations will fix in a population each generation. However, the average number of mutations which will fix in a population each generation is, from Definition 1, the average rate of substitution per generation, k, and the result follows.

■

The remarkable consequences of Theorem 1 became the basis for neutral model of molecular evolution. Kimura and Ohta (1971) have argued that the rate of amino acid substitution per year is remarkably constant among vertebrate lineages for each protein they have studied. This has led to the concept of a molecular clock, which is used in author's study described in chapter 5, section 5.3.

Kimura and Ohta (1971) have measured the average rate of substitution among the proteins examined to be $k_s = 1.6 \times 10^{-9}$ amino acid substitutions per amino acid site per year. If substitutions were neutral, then from equation (1) it follows that the average neutral mutation rate, μ_s, should be 1.6×10^{-9} amino acid mutations per amino acid site per year, which is strikingly close to nucleotide mutation rates as measured in laboratories (Gillespie 1998). Kimura and Ohta (1971) found this result as a strong argument for the neutral model. They argued that, as typical protein heterozygosities are around 0.1, it follows from Theorem 3.3:1, equation (3.3:2) that $4N\mu$ must be approximately 0.1 as well.

Since a typical protein is about 300 amino acids long, and about 30% of the variation is detected by the electrophoresis in experiments, which Kimura and Ohta (1971) relied on, thus, the mutation rate to electrophoretically detectable variation for the entire protein is $\mu = \mu_s \times 300 \times 0.3 = 1.6 \times 10^{-9} \times 300 \times 0.3 = 1.44 \times 10^{-7}$. This result was considered as a slight overestimate, and they used $\mu = 10^{-7}$ for the remainder of their investigation.

Assuming that mice have on average two generations per year, they estimated the effective population size of mice from the formula $4N\mu = 4N \times 10^{-7}/2 \approx 0.1$, which gives $N \approx 5 \times 10^5$. Note that this is long-term effective population size, which according to equation (3.2:22) in Theorem 3.2:1 is the harmonic mean of short-term effective population sizes. Similarly, the neutral model applied to electrophoretic data imply that the long-term effective population size for humans is about 1.25×10^4.

Since the time of Kimura's famous book (Kimura 1983), the neutral theory of molecular evolution has become the dominant explanation for most of protein and DNA evolution. The theory has encountered some problems, particularly with regard to protein evolution where the generation-length effect was not observed. It should be noticed however, that the effect of generation-length has been however detected in non-coding DNA sequences, which are considered neutral (Gillespie 1998). Therefore, it is a common view nowadays, that most amino acid substitutions are not neutral but are slightly deleterious, and thus are less frequent than predicted under the strictly neutral model. It also implies that the heterozygosity observed in population is smaller than it would have been under neutral model, what explains (in addition to the effect of harmonic mean) the unexpectedly small long-term effective population sizes of many species, including humans, estimated based on protein polymorphism data.

Despite these difficulties with explanation of protein variation, the neutral theory of evolution served as a theoretical model for development of statistical neutrality tests. With advent of these tests, the search for signatures of natural selection operating at the molecular level has become more and more important (this problem is discussed in section 4.3). It is so because neutral theory of evolution at molecular level, proposed by Kimura and Ohta (1971), does not deny the existence of selection observed at that level. It only states that the majority of observed genetic variation is caused by random fluctuation of allele frequencies in finite populations (effect of genetic drift – see section 3.2) and by selectively neutral mutations (see section 3.3).

If majority of mutations have been claimed to be neutral, then the next step should be to search for those which are not neutral. Therefore, as mentioned above, several statistical tests, called neutrality tests, have been developed (they are described in detail in section 4.2) and the neutral theory of evolution has been used as a null hypothesis for them. A statistically significant departure from this model can be therefore treated as a signature of natural selection operating in a gene under consideration.

Unfortunately, other reasons for departure from the neutral model are also possible and they also account for statistically significant signals in neutrality tests. These reasons include expansion of the population (problem of discovery of population expansion is discussed in

section 5.2) and geographical substructure of population with limited migration among demes. Also recombination accounts for incorrect testing of the natural selection often suppressing the positive test signals even if the selection was present. Moreover, these effects affect various tests with different strength, resulting in an interpretation puzzle instead of clear indication in the favor of the natural selection or against it.

Aforementioned difficulty in the interpretation of a battery of tests is the start point for application of the author's multi-null-hypotheses (MNH) method (see section 4.3.2). The author has co-developed multi-null hypotheses methodology (partially published in Cyran et al. 2004, and lately further improved) capable for the reliable interpretation of the test outcomes in the context of natural selection. However, since the method requires modified null hypotheses, the critical values of the tests are unknown and the huge amount of computer simulations must be carried out for estimation of these values. Therefore, the AI-based methodology, including the author's QDRSA described in section 2.3.3, was proposed as an efficient and fast solution. The application of AI-based methods in the problem of the search for natural selection is presented in section 4.3.3.

4.2. Neutrality tests

As mentioned in section 4.1, testing for natural selection operating at molecular level has become one of the important issues in contemporary bioinformatics. Such research relies on development of neutrality tests, which are statistics that can be used against null hypotheses based on predictions of the neutral model of evolution. These tests can be often used in search for signatures of natural selection in genes, as presented in section 4.3

There exist two general types of tests of natural selection at molecular level. The first type can be applied when the data consists of entire or partial coding sequences of a gene. Then, the comparison of frequencies of silent substitutions at the third codon position to the frequencies of substitutions on the first and second position provides a handle to measure selective pressure. This approach was used in study leading to detection of perhaps the most spectacular example of natural selection found in the ASPM locus, a major contributor to brain size regulation in primates (for more information about evolution of ASPM see Evans et al. 2004, Zhang 2003).

In many cases, however we have to deal with another type of data, which consists of sequences that are not only non-coding, but also composed of nucleotides located at a considerable distance from each other. In such cases, a model for neutral evolution of the sequence has to be determined and then its predictions compared to data. Usually, this model is some modification of the Wright-Fisher model of genetic drift with mutation (Hartl and

Clark 1997, Jobling et al. 2004). The significant departure from predictions under neutrality (which serves as the null hypothesis) may provide evidence for selection (the desirable alternative hypothesis). However, there exist other alternatives, which may cause departures from the null, and be confused with selection. Examples include population substructure and past change in population size as reviewed by Nielsen (2001). Therefore, one common way to deal with this problem is to frequently apply a number of tests, each one sensitive to different combination of factors, and compare the results. The substructure for example can be approached by considering data from different subpopulations separately or by comparison of the test results among loci. Another approach presented in the section 4.3.2 is based on the formulation of null hypotheses assuming population substructure. In this way, if proper critical values of the test are determined, the influence of substructure will not cause false positive test results.

Each analysis of SNP data leading to the detection of natural selection operating at some loci, when applied to human population, has to take into consideration the alternative departures from neutrality that can produce data resulting in similar test outcomes. These alternatives feasible from the point of view of human population evolution are population growth and geographic substructure with migration. In section 4.3 it will be shown how to deal with this problem, by the analysis of a battery of statistical tests giving indication about the age of the predominant mutations, and how this information can be used to exclude not desirable alternatives. In this section these neutrality tests will be defined.

Tests which give the indication about the age of alleles, being in excess compared to the amount predicted under neutral evolution model, are based on the difference between different estimates of composite parameter $\theta = 4N\mu$ (N indicates the effective population size and μ is the mutation rate per nucleotide per generation). Such tests are Fu's tests belonging to the class $F'(r, r')$ (Fu 1997):

$$F'(r,r') = \frac{L'(r) - L'(r')}{\sqrt{Var\left[L'(r) - L'(r')\right]}}, \tag{4.2:1}$$

where L' are estimates of composite parameter θ in the form of linear functions of the η_i (the numbers of segregating sites of type i, where $i = 1, 2, \ldots, \lfloor n/2 \rfloor$ and n is the sample size). The parameter of function L' denotes more (for larger values) or less (for smaller values) substantial influence of rare alleles on the estimation of θ. Therefore, $\hat{\theta}_\pi = L'(0)$ is less influenced by rare alleles than $\hat{\theta}_W = L'(1)$. The defined above class covers many known tests like: Tajima (1989) test T (for uniformity, we follow the nomenclature of Fu (1997), Wall (1999), and some other papers, although originally Tajima's test was named D), Fu and Li's (1993) test D^* or Fu and Li's (1993) test F^*.

Definition 4.2:1 (Tajima test *T*, after Tajima 1989)

The Tajima test *T* is defined as the normalized difference between the estimates of composite parameter $\theta = 4N\mu$ based on the average genetic distance $\hat{\theta}_\pi$ and the number of segregating sites:

$$T = \frac{\hat{\theta}_\pi - \hat{\theta}_W}{\sqrt{Var\left(\hat{\theta}_\pi - \hat{\theta}_W\right)}}. \tag{4.2:2}$$

■

Tajima *T* test, which is the most widely used neutrality test (McVean 2002), is equivalent to *F*'(0,1). Other tests of *F'*(*r*, *r*') class include Fu and Li's test *D** (*D** = *F'*(1,∞) and therefore the test is sensitive to existence of very rare alleles) and Fu and Li's test *F**. Since *F** = *F'*(0, ∞) it should have the power for detecting the excess of very rare alleles, presumably with greater power than *D** because of a more extreme value of the first parameter in function *F'*.

Definition 4.2:2 (Fu and Li tests, after Fu and Li 1993)

Fu and Li tests *D** and *F** are defined as

$$D^* = \frac{\frac{n}{n-1}\eta - \eta_s \sum_{i=1}^{n-1} \frac{1}{i}}{\sqrt{u_{D^*}\eta + v_{D^*}\eta^2}}, \qquad F^* = \frac{\hat{\theta}_\pi - \frac{n-1}{n}\eta_s}{\sqrt{u_{F^*}\eta + v_{F^*}\eta^2}}, \tag{4.2:3}$$

where η is the total number of mutations that occurred in the entire genealogy of *n* genes, and η_s is the number of singletons, *i.e.* nucleotides that appear only once at the site among the sequences in the sample. For mathematical definitions of coefficients u_{D^*}, v_{D^*}, u_{F^*} and v_{F^*} (being complicated functions of the parameter *n* only) see Fu and Li (1993).

■

Another category of tests is based on the estimates of probabilities of having no more or no less than the observed number *k* of haplotypes in a sample of *n* sequences, assuming neutrality and lack of intra-locus recombination. Into this category fall the Strobeck's test *S* and the Fu's test F_s, which are defined below.

Definition 4.2:3 (Strobeck's test *S*, after Cyran, Polańska, and Kimmel 2004)

The Strobeck's test *S* is defined as the estimate of the probability of having no more haplotypes in a sample, and it is given by

$$S = \sum_{i=1}^{k} \frac{\left|S_n^i\right| \hat{\theta}_\pi^i}{S_n\left(\hat{\theta}_\pi\right)} \tag{4.2:4}$$

where: $S_n(\hat{\theta}_\pi)$ denotes the generating function of the Stirling numbers of the first kind S_n^i, *i.e.*

$S_n(\hat{\theta}_\pi) = \sum_{i=0}^{n} S_n^i \hat{\theta}_\pi^i = \hat{\theta}_\pi(\hat{\theta}_\pi+1)\ldots(\hat{\theta}_\pi + n - 1)$.

Definition 4.2:4 (Fu's test F_s, after Cyran, Polańska, and Kimmel 2004)

The Fu's test F_s is given by:

$$F_s = \ln\left(\frac{S'}{1-S'}\right), \tag{4.2:5}$$

where *S*' is the estimate of the probability of having no less than observed number *k* of haplotypes in a sample of *n* sequences. Therefore (compare with (4) for similarities) it is given by:

$$S' = \sum_{i=k}^{n} \frac{|S_i| \hat{\theta}_\pi^i}{S_n(\hat{\theta}_\pi)}. \tag{4.2:6}$$

In the framework of the infinite allele model (IAM), the SNP haplotypes are treated as new variants (mutants) of a SNP sequence. Ewens Sampling Formula, derived under neutrality and no recombination, provides expected frequencies of haplotypes existing in a given number of copies (Hartl and Clark 1997). Therefore, it serves as a convenient reference to test deviations from neutrality. It is used, for example, in the Strobeck's test (see above). However, it is even more convenient to use coalescent simulations based on the IAM, to compute a large sample of simulated distributions of variants. The value of composite parameter θ is estimated from the haplotype sample, using the IAM-based expression for the total number *K* of variants in the sample of *n* sequences:

$$K(\theta) = \sum_{i-1}^{n} \frac{\theta}{\theta + i - 1} \tag{4.2:7}$$

and comparing it to the observed number of different haplotypes. Simulated distributions of variants are compared to the observed frequencies. Technically, to facilitate visual comparison, empirical and simulated cumulative counts *A*(*j*) of haplotype variants existing in no more than *j* copies in the sample of *n* sequences (*j* = 1, ..., *n*) are compared. In addition, both the horizontal axis (the number *j* of copies of a variant) and the vertical axis (cumulative count *A*(*j*) of variants existing in *j* copies) are standardized to the unit interval, by dividing by *n* and *K*, respectively. Resulting graphs allow a visual comparison of the empirical distribution of variants (thick line) with multiple simulated distributions (thin lines), as presented for illustration in Fig. 1 and Fig 2 for actual SNP data.

Definition 4.2:5 (Kelly's test, after Kelly 1997)

The Kelly's test Z_n is defined as the average (over all pairs *i, j* of *K* segregating sites) of the squared correlation of allelic identity between sites *i* and *j*

$$Z_{nS} = \frac{2}{K(K-1)} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} \delta_{ij.} \tag{4.2:8}$$

■

If the goal is eventually to find the type of selection, at first one should exclude Kelly's (1997) Z_{nS} test, as it produces similar, inflated, patterns both for selective sweeps with recombination and for balancing selection. Note, however, that it is valuable to apply the Z_{nS} test after one of these possibilities has been excluded based on results of the tests given by Definitions 1 – 4. It is so because this test is reported to have a big power, and can verify previously obtained results.

The squared correlation of allelic identity δ_{ij} is a standardized (that is ranging from 0 to 1) measure of linkage disequilibrium D_{ij} between loci *i* and *j*. It is given by:

$$\delta_{ij} = \frac{D_{ij}^2}{p_i(1-p_i)p_j(1-p_j)}. \tag{4.2:9}$$

In above formula $D_{ij} = p_{ij} - p_i p_j$, where p_i and p_j are frequencies of mutant alleles at loci *i* and *j* respectively, whereas p_{ij} is the frequency of sequences that have mutant alleles at both loci.

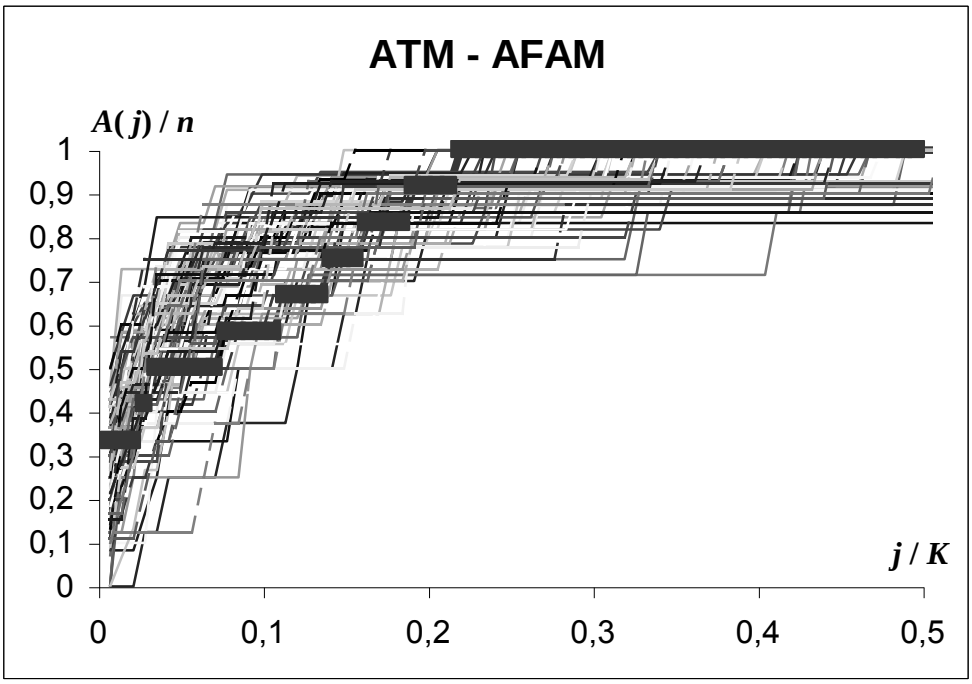


Fig. 4.2:1. Graphical depiction of nonneutrality at ATM obtained from simulations
Rys. 4.2:1. Graficzna prezentacja braku neutralności w ATM na podstawie symulacji

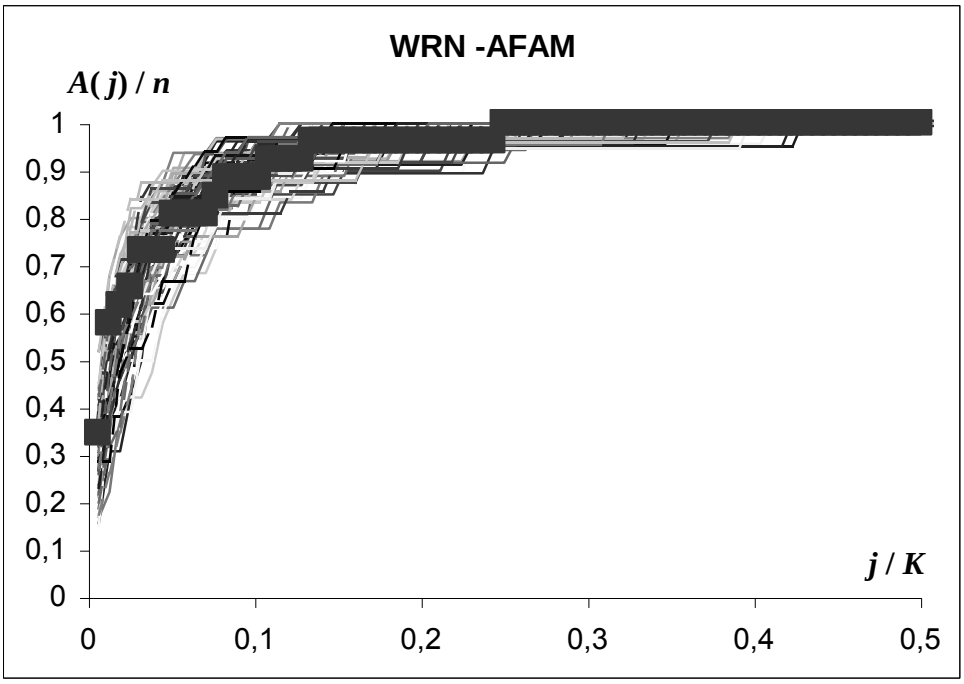


Fig. 4.2:2. Graphical depiction of neutrality at WRN locus obtained from simulations
Rys. 4.2:2. Graficzna prezentacja neutralności w lokusie WRN na podstawie symulacji

The last tests presented here are Wall's (1999) tests *B* and *Q*.

Definition 4.2:6 (Wall's test *B*, after Wall 1999)

Statistic *B* is defined as the normalized number *B'* of pairs of adjacent congruent (*i.e.* inducing identical partitions of the set of haplotypes) segregating sites. To be normalized, *B'* is divided by the total number $(K-1)$ of pairs of adjacent segregating sites:

$$B = \frac{B'}{K-1}. \tag{4.2:10}$$

■

Definition 4.2:7 (Wall's test *Q*, after Wall 1999)

Let us indicate by A the set of all distinct partitions induced by pairs of adjacent congruent segregating sites. Then the statistic *Q* is defined as

$$Q = \frac{B + card(A)}{K} \tag{4.2:11}$$

where $card(A)$ is the number of elements in *A*.

■

Note, that the power of a test *Q* becomes less sensitive to the recombination as compared to the test *B*, because the decrease of *B* is compensated by the increase of *card* (A) in a presence of recombination.

The careful analysis of the mentioned battery of tests if applied to many loci and for many subpopulations can give the answer about the presence of natural selection at some of them. Theoretically, population growth and population substructure effects should be identical (or in the presence of recombination, similar) for all loci. Any large difference in test outcomes among loci is a signal that some specific to some loci reason is probably non-negligible cause of detected departures. Also analysis of relatively genetically pure subpopulation can reveal that the cause of departure from neutrality is not the substructure (since in such subpopulation it is not of the main importance). Yet in practice it is not easy to obtain a sample from genetically pure population since admixtures accumulated over long time are of different intensity in main human subpopulations (Budowle and Chakraborty 2001, Budowle *et al*. 2001, Chakraborty 1986).

Because of mentioned reasons it may be helpful to employ more sophisticated null hypotheses. Certainly they should assume neutrality (which is subject to be rejected by the test result) but on contrary to standard null hypotheses they can incorporate more feasible population models. The degree to which they can imitate the real history of human population depends on our knowledge about this history (still very incomplete in long term) but they always should be formulated to be conservative with respect to feasible population history scenarios (in order to prevent too many false positives). The exact meaning of being conservative in this aspect is dependent on the actual data. Therefore it is always desirable to perform the battery of mentioned above tests with standard null hypotheses and infer based on them whether departure from neutral model is in direction of excess of old or young mutations. The excess of young mutations is characteristic for positive selective sweep or for slightly deleterious mutations, whereas the excess of old mutations is observed in loci under balancing selection pressure. Since the population expansion is also the cause of many young mutations, therefore modified null hypothesis assuming growth would be more conservative than standard in search for selective sweep, but less conservative than standard in search for balancing selection. On the other hand the effect of population substructure shifts the excess of alleles in opposite direction as compared to population growth.

4.3. Search for selection at molecular level – case study

After presenting neutrality tests in the section 4.2, this section describes using them in search for signatures of natural selection in SNP haplotypes taken from the intronic regions of four genes implicated in human familial cancers: ataxia telangiectasia mutated (ATM), human helicase RECQL, Bloom's syndrome (BLM) and Werner's syndrome (WRN). An attempt to explain the origin of human-chimpanzee trans-specific polymorphism discovered

in one SNP of ATM is also given. The sample is composed of about 600 chromosomes, derived from residents of Houston, TX (USA), representing major ethnic backgrounds: Caucasian, African-American, Asian-American and Hispanic. Deviations from neutrality may be obscured by presence of recombination, substructure and changes of population size.

To investigate these effects on data presented in section 4.3.1 there was applied a novel author's methodology based on conservative modifications of null hypotheses to invoke effects of population growth, population substructure and recombination (section 4.3.2). Additionally fast screening procedure based on artificial intelligence methods is given (section 4.3.3). In two loci (ATM and RECQL) there were found signatures of balancing selection preserving excess of older mutations. In the case of ATM, balancing selection supports hypothesis that origin of a bi-allelic polymorphism, shared by humans and chimpanzees, predated speciation. The variability pattern observed in BLM and WRN can be explained within neutral model.

4.3.1. Data: Single-nucleotide polymorphisms in four gene regions

There is analyzed a total of 45 Single Nucleotide Polymorphisms (SNPs) located on intronic and other non-coding sequences of the ATM gene, and three human helicases BLM, WRN, and RECQL. Tables 1 – 4 inform about names, positions and variations of the analyzed SNPs.

Table 4.3:1
Name, positions with respect to the beginning of the sequence having accession number given in the first row, and variations of the analyzed SNPs within ATM locus

ATM [U82828]		
Prior to 5'UTR t-a	10182	T→A
IVS8-356t-c	34293	T→C
IVS19-1276a-g	57469	T→C
IVS21-77t-c	60136	T→C
IVS34+754g-a	85811	C→T
IVS46-257a-c	112721	A→C
IVS55+186c-t	121819	C→T
IVS57+3570t-c	127195	T→C
IVS58+997g-a	132032	C→T
IVS61-55t-c	142611	T→C
IVS62+60g-a	142789	C→T
IVS62+424g-a	143153	C→T
IVS62-973a-c	151964	T→C
IVS62-694c-a	152243	C→A

Table 4.3:2
Name, positions with respect to the beginning of the sequence having accession number given in the first row, and variations of the analyzed SNPs within RECQL locus

RECQL [AC006559]		
IVS1-89964t-g	10998	A→C
IVS1-42581t-c	58381	T→C
IVS1-30638g-c	70324	G→C
IVS1-30329g-t	70633	G→T
IVS1-24228g-a	76734	G→A
IVS1-24159c-t	76803	G→A
IVS1-7216a-g	93746	T→C
IVS1-7166g-a	93796	G→A
IVS10-1078g-a	113771	G→A
IVS15+19546t-c	152798	T→C
IVS15+33444t-c	166696	T→C

Table 4.3:3
Name, positions with respect to the beginning of the sequence having accession number given in the first row, and variations of the analyzed SNPs within WRN locus

WRN [AF181896]		
IVS1-8213g-a	6114	G→A
IVS4+176a-g	45121	T→C
IVS19-3173t-c	88968	T→C
IVS19-3145t-a	88996	T→A
IVS24-191c-t	111606	G→A
IVS32+845c-t	135048	G→A
IVS32+859g-t	135062	G→T
IVS34-628t-g	145865	A→C
IVS35+4302t-c	157465	T→C
IVS35+11737g-c	164900	G→C
IVS53+30673c-t	183836	G→A
IVS35+30764c-a	183927	G→T

The ATM gene located in human chromosomal region 11q22-q23 (Fig. 1a) spans 184 kb of genomic DNA (Bonnen *et al.* 2000) and contains 66 exons (Uziel *et al.* 1996). Yu *et al.* (1997) determined the intron-exon structure of the WRN locus spanning 186 kb at 8p12-p11.2 (Fig. 1b) and found 35 exons, with the coding sequence beginning in the second exon. RECQL is composed of 15 exons, located at 12p12-p11 (Fig. 1c) and spans 180 kb, whereas

BLM mapped to 15q26.1 (Fig. 1d) has 22 exons and spans 154 kb (Trikka *et al.* 2002). The regions used for SNP scanning were in total 13.5 kb long for ATM (Bonnen *et al.* 2000) and covered between 15% and 20% of the three helicases (Trikka *et al.* 2002).

Table 4.3:4
Name, positions with respect to the beginning of the sequence having accession number given in the first row, and variations of the analyzed SNPs within BLM locus

BLM [AC002312]		
IVS1-20561t-c	21812	T→C
IVS1-20290g-a	22083	G→A
IVS17-425a-g	122762	T→C
IVS17-345c-g	122842	C→G
IVS22-2082c-a	136931	G→T
IVS22+3336c-g	142615	C→G
IVS22+3401a-c	142680	A→C
IVS22+9303c-t	148582	G→A

Detailed data on primer sequences, PCR conditions and product sizes for each of the polymorphic sites, as well as the ASO hybridization sequences and wash conditions for each SNP variant for the ATM gene can be found in Bonnen *et al.* (2000) and for the BLM, WRN and RECQL genes in Trikka *et al.* (2002). Blood samples were collected from residents of Houston, TX, belonging to four major ethnic groups: Caucasians, African-Americans, Hispanics, and Asians (Table 5).

Table 4.3:5
Number of chromosomes in each ethnicity/locus group

Ethnic group	**BLM**	**WRN**	**RECQL**	**ATM**
African-Americans	146	154	156	142
Caucasians	152	158	156	154
Hispanics	144	150	152	146
Asians	78	78	74	78

The screening protocol used for discovery of SNPs most probably has missed less than 10% of SNPs actually present in samples used for this purpose (Trikka *et al.* 2002). Haplotypes were inferred and their frequencies were estimated using the Expectation-Maximization (EM) algorithm (Dempster, Laird, and Rubin 1977, Excoffier and Slatkin 1995, Polańska 2003). The estimated recombination rates $C = 4N_e c$ (Hudson 1987) are shown in Table 6 and sequences of great apes (Bonnen *et al.* 2000, Trikka *et al.* 2002) corresponding to human SNPs are shown in Table 7.

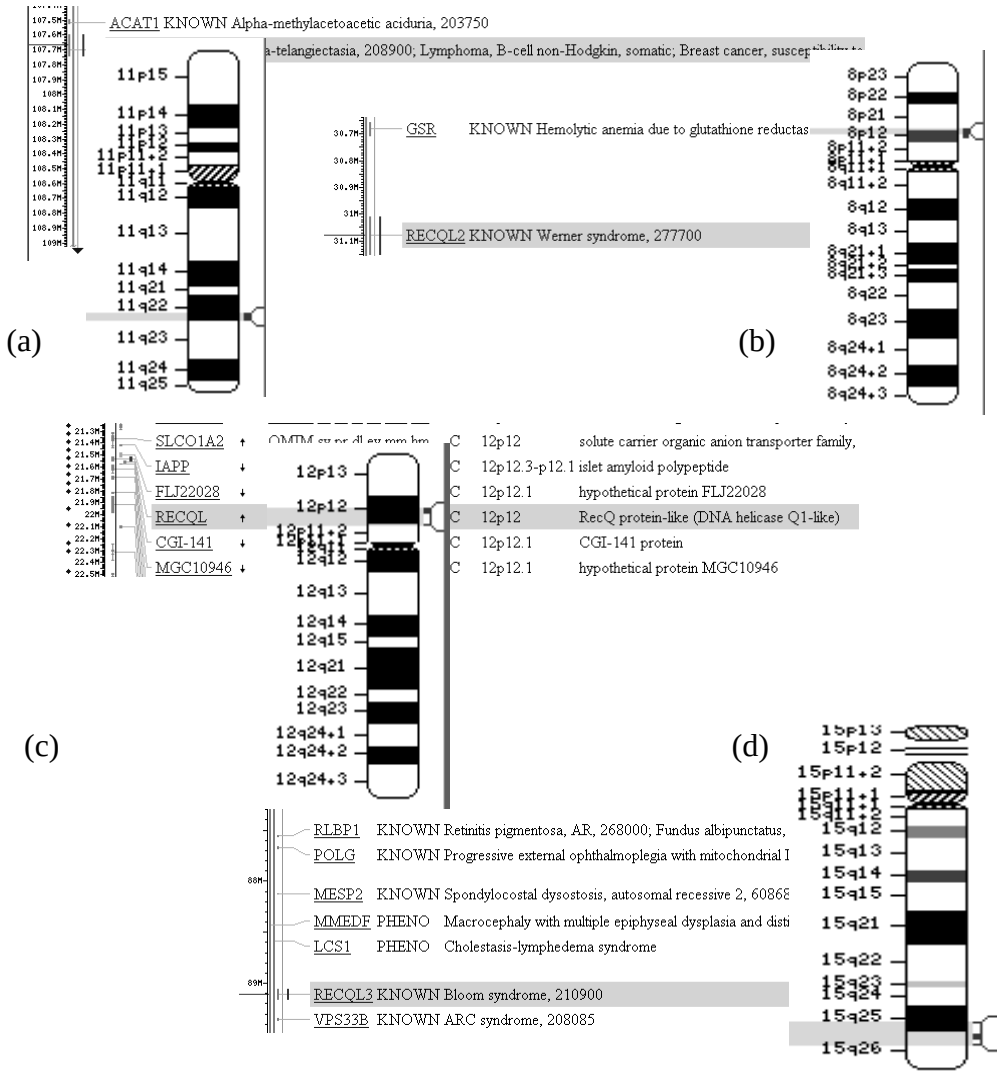


Fig. 4.3:1. Four genes under study: (a) ATM, (b) WRN, (c) RECQL, and (d) BLM
Rys. 4.3:1. Cztery rozważane geny (a) ATM, (b) WRN, (c) RECQL, and (d) BLM

Table 4.3:6

Estimated values of recombination rate $C = 4N_e c$ per gene

Recombination C [per gene]	AfAm	Caucasian	Asian	Hispanic	Global
ATM	5.6	2.6	0.4	1.7	3.3
RecQL	9.2	3.5	0.7	4.2	5.1
WRN	41.6	34.8	12	16.4	28
BLM	32.5	16.8	16	23.6	29.2

Table 4.3:7

Sequences of great apes corresponding to human SNPs analyzed

	ATM	RECQL	WRN	BLM
Pan troglodyte 1	TCTTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
	TCTTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
Pan troglodyte 2	TCTTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
	TCTTTACTCTCTTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
Pan paniscus	TCTTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
	TCTTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
Gorilla gorilla 1	TATTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
	TATTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
Gorilla gorilla 2	TATTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
	TATTTACTCTCCTC	ATGGAGTGGTT	GTTTGGGATGGG	CGCCGCAG
Gorilla gorilla graueri 1	TATTTACTCTCCTC			
	TATTTACTCTCCTC			

4.3.2. Multi-null-hypotheses method

Until recently, demonstrations of natural selection at the molecular level in the human genome were not so numerous. However, by now, there is a number of examples (Bamshad *et al*. 2002, Gilad *et al.* 2002, Toomajian and Kreitman 2002, Wooding *et al.* 2002), with perhaps the most spectacular being the ASPM locus, a major contributor to brain size regulation in primates (Zhang 2003, Evans *et al.* 2004). Usually, the model used for detection of selection is the Wright-Fisher model of genetic drift with mutation. Significant departure from predictions under the null hypothesis of neutrality may provide evidence for an alternative hypothesis of selection. However, there exist other alternatives, which may cause departures from the null, mimicking the effect of natural selection. Among these, the most important are population substructure and past change of population size (Nielsen 2001). These influences may be difficult to disentangle from effects of selection. In this section there is described the author's approach based on applying a series of nested null hypotheses, instead of just one. Comparison of test outcomes against these nulls will, arguably, help eliminate genetic and/or population-related factors other than selection as causes of departures from strict neutrality.

In a series of papers (for example Bonnen *et al.* 2000, Bonnen *et al.* 2002, Trikka *et al.* 2002) scientists from Houston genetic centers were investigating SNP haplotypes at four genes: ataxia telangiectasia mutated (ATM), human helicase RECQL, Bloom's syndrome (BLM), and Werner's syndrome (WRN). Since these genes are also implicated in human

familial cancers and impaired DNA repair, they could be potentially subject to natural selection.

ATM gene product is a member of a family of large proteins implicated in regulation of the cell cycle and response to DNA damage. Predominant abnormalities in this gene, which exhibits a remarkable diversity, involve point mutations or small rearrangements leading to splicing mutations (Teraoka *et al.* 1999). Li and Swift (2000) determined that patients heterozygous for splice site mutations have significantly longer survival than those homozygous for single truncating mutations. Some of the ATM mutations are responsible for ataxia telangiectasia, a recessive pleiotropic disorder, clinically characterized by cerebellar ataxia, oculcutaneous telangiectasia, immunodeficiency, sensitivity to radiomimetic agents, and predisposition to cancer.

In mentioned above work (Bonnen *et al.* 2000, 2002) the analysis of haplotypes revealed reduced recombination and extensive linkage disequilibrium at the ATM locus. Due to this, association studies using ATM haplotypes have a significant potential for detection of genetic backgrounds that contribute to disease. By comparison of detected SNPs with corresponding sequences of great apes our group discovered a bi-allelic polymorphism shared by humans and chimpanzees (Bonnen *et al.* 2000). Perhaps this polymorphism arose independently in the two species, but if it were the consequence of polymorphism present in a common ancestor of humans and chimpanzees, the finding would imply the existence of very old mutations in ATM. The latter hypothesis is consistent only with overdominance at the ATM locus because only such form of selection can preserve mutations for an almost arbitrarily long time (Slatkin and Rannala 2000). However, until author's works no tests of this hypothesis were performed.

The remaining three genes analyzed are human DNA helicases. All polypeptides encoded by these genes share a central region of seven helicase domains (Siitonen *et al*. 2003). They are involved in many aspects of DNA metabolism, including transcription, accurate chromosomal segregation, recombination, and repair. Helicase-dependent DNA repair include mismatch repair, nucleotide excision repair, and direct repair. Since genomes are subject to damage by chemical and physical agents in the environment, as well as by free radicals, endogenously generated alkylating agents or replication errors, the genetically determined effectiveness of repair is one of the important factors deciding about the fitness of corresponding phenotype.

Bloom and Werner syndromes, being similarly as ataxia telangiectasia rare autosomal recessive disorders, have overlapping clinical features, of which high predisposition to malignancies is the most remarkable (Siitonen *et al.* 2003). WRN plays an additional role in preventing premature aging via a mechanism suggested to be common for eukaryotes (Sinclair *et al.* 1997) and is involved in exonuclease activity (Huang *et al.* 1998). It has BLM-

binding regions containing N-terminal exonuclease domain with activity inhibited by BLM binding. At the same time, the WRN helicase activity is not affected by BLM binding (Von Kobbe *et al.* 2002). Cells in Bloom syndrome exhibit hypermutability including hyperrecombinality between sister chromatids and homologous chromosomes (Yusa *et al.* 2004). Karow *et al.* (2000) emphasizes the role of BLM as an antirecombinase for suppresion of tumorigenesis. Wu and Hickson (2003) have proposed a similar mechanistic explanation of BLM-based tumorigenesis suppression. BLM-catalized dissolution of double Holiday junctions prevents sister chromatid exchange and through suppression of ectopic recombination and crossing-over between homologous chromosomes BLM product prevents loss of heterozygosity. Adams *et al.* (2003) concluded that BLM maintains genomic stability by promoting efficient repair DNA synthesis and thereby prevents double-strand break repair by less precise pathways.

Interestingly, Ellis *et al.* (1994) have determined that a 6-bp ATCTGA deletion and 7-bp TAGATTC insertion at nucleotide 2281 of BLM cDNA, is a mutation inherited from a founder of Ashkenazi Jewish population and nearly all Ashkenazi Jews with Bloom syndrome inherit this mutation, named blm^{Ash}, identical by descent from this common ancestor. Cells derived from individuals suffering from any the two syndromes show significant levels of genomic instability caused by the increased level of chromosomal aberrations (Yamagata *et al.* 1998), however RECQL has not been related to any disease and its functions, other than DNA unwinding, remain unknown. Geneticists from Houston (Trikka *et al.* 2002) performed detailed linkage disequilibrium and recombination analysis for these helicases with results not as extreme as for the ATM. For the BLM we confirmed the founder haplotype of Ashkenazi Jews homozygous for blm^{Ash}.

The range of functions crucial for survival enumerated above as well as the characteristic patterns of polymorphism present in our samples suggest that these genes may be under selective forces possible for detection. The simplest directional deleterious selection that may be postulated is unlikely due to existence of old mutations in all loci. More feasible is a form of balancing selection. The current section tackles the problem of identification of selection and presents a methodology based on incorporating demography into null hypotheses.

To detect departures from the neutral model, the following statistics described in detail in section 4.2 were used: Tajima's (1989) T (for uniformity, the nomenclature of Fu (1997) and Wall (1999) is followed), Fu and Li's (1993) F^*, Kelly's (1997) Z_{nS} and Wall's (1999) Q. The choice of above tests was dictated by: (a) the type of data at disposal, and (b) by the proposed methodology of verification whether a detected departure from the neutral expectation can be considered to be a result of a given type of selection operating at the locus: Issues, assigned above as (a) and (b) are discussed in more detail in what follows.

a) Since the SNPs analyzed come from intronic regions of the target genes, it was not possible to use McDonald-Kreitman (1991) type tests based on the differences in ratios of nonsynonymous and synonymous mutation rates within and between species (resulting in polymorphism and divergence, respectively), although they are reported to be very powerful in detection of selection and not dependent on population demographic effects (Nielsen, 2001). Similar reasons excluded the application of Akashi's (1995) test, as well as Nielsen and Weinreich's (1999) test, in which the ages of nonsynonymous and synonymous mutations are estimated and compared with predictions of the neutral model. Hudson, Kreitman and Aguade's (1987) HKA test was not used due to lack of chimpanzee sequences for all introns containing our SNPs. The test using interspecific divergence rate calculated from only a few introns that could be obtained using BLAST search of the databases of the Chimp Sequencing Project, was considered to be potentially biased.

b) Natural selection is not the only genetic force causing departures from predictions of the neutral Wright-Fisher model in the usual form, *i.e.* assuming a panmicting population and constancy of the population size. Since none of these assumptions strictly holds for actual human demography, there is a proposition to incorporate demographic effects into null hypotheses. Then, the departure from the modified nulls could be considered as caused by selection. One of the delicate points of this approach is that scientists only know a general outline of the past human demography. Nevertheless, it is possible to assume demography that is more realistic than that assumed in the classical Wright-Fisher model, and at the same time, which is conservative. Conservative means that it is more difficult to reject the null with this assumed demography than it would have been with the actual unknown demography. It implies that we have to use conservative parameter values for growth and migration rates in expanding and sub-structured human population. These parameter values are different for different types of selection. This is why it is so crucial to know, before proposing modified null hypotheses, whether the genealogy implied by data is similar to that caused by (*i*) growth, deleterious selection or positive selective sweeps, or (*ii*) shrinkage, substructure or balancing selection. As it was presented in greater detail in the section 4.2, tests which can reliably assign the pattern of departure to (*i*) or (*ii*), are these belonging to Fu's (1997) $F'(r,r')$ class. Tajima's T and Fu's F^* are two the most extreme cases of such tests: $F'(0, 1)$ and $F'(0, \infty)$ respectively. The first relies on estimates of $\theta = 4N\mu$ based on the average number of nucleotide differences and on the number of segregating sites, the second compares the number of mutations located on external and internal branches of genealogy. Similar idea of comparison of the lengths of old and recent branches of genealogy is incorporated in Kelly's Z_{nS} statistic based on the

average linkage disequilibrium at the locus. However, this latter produces similar, inflated, patterns both for selective sweeps with recombination and for balancing selection. Also, Wall's *W* and *Q* tests based on the number of adjacent congruent segregating sites employ similar principle. The latter pair of tests is reported (Wall 1999) to be especially well designed for detection of balancing selection, which may be suspected to operate on genes associated with disease and presenting a polymorphism with the excess of old mutations (test *Q* is preferred over *W* if recombination is present).

In order to exclude genetic forces other than selection as sources of significant test outcomes, we applied four different null hypotheses:

- H_{00}, panmictic population, with population size constant in time,
- H_{01}, panmictic population, with population size increasing exponentially 10 times over the period of 5,000 human generations, to achieve present effective population size N_{end} = 100,000.
- H_{02}, sub structured population, growing like in $H_{01,}$ composed of 4 demes with a split 5,000 generations ago and between-deme migration rate $m \times N_{\text{end}} = 100$.
- H_{03}, demography like in H_{02}, but with recombination with estimated intensities (Table 6).

The influence of genetic forces assumed in the null hypotheses on site frequency spectra of the ATM gene for African Americans, predicted under selective neutrality, is presented in Figure 2.

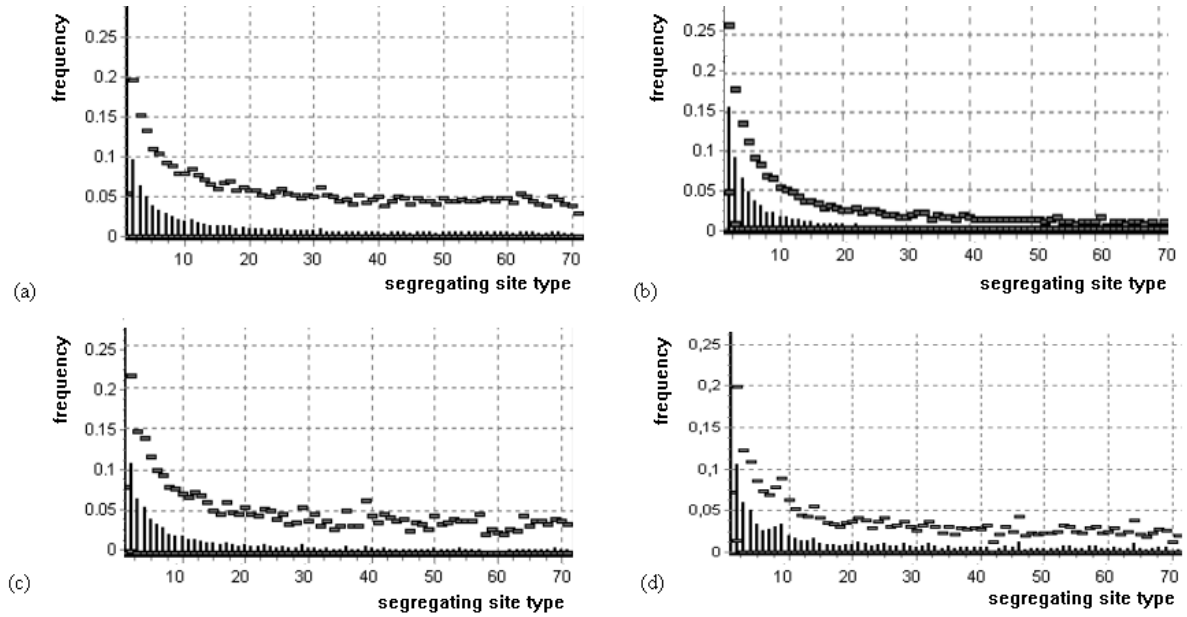


Fig. 4.3:2. The illustration of the influence of null hypothesis on expected frequencies of segregating sites of types: 1 to $\lfloor n/2 \rfloor$

Rys. 4.3:2. Ilustracja wpływu hipotezy zerowej na oczekiwane częstości pozycji segregujących typoów: 1 do $\lfloor n/2 \rfloor$

The segregating site is said to be of the type *i* if it has *i* and *n-i* variants in a sample, therefore, the less frequent the segregating site is, the closer to one is its type (reaching one for singletons). Charts in Fig. 1 present simulated frequencies of a sample composed of $n = 142$ sequences 13.5 kbp long, conditioned on 13 segregating sites (corresponding to ATM sequence for AfAm population) assuming selective neutrality under null hypotheses (a) H_{00}, (b) H_{01}, (c) H_{02} and (d) H_{03}. Observe excess of rare segregating sites and reduction of frequent segregating sites under H_{01} compared to H_{00}. Such reduction is characteristic for samples corresponding to all considered genes and populations (results not shown). The H_{02} and H_{03} result in slight excess of rare segregating sites over H_{00}. Since the neutral site frequency spectrum changes for various null hypotheses, so should the critical values of tests based on the shape of such spectra (for example T or F^*). Horizontal axis denotes the type of the segregating site, while vertical axis shows the relative frequency of the site of a given type. In the charts, vertical bars indicate the average frequencies over all simulations, whereas horizontal upper and lower bars indicate maximum and minimum values of these frequencies, respectively. Note that horizontal lower bars, for all types of segregating sites except the rarest, indicate frequency zero, and therefore are hardly visible.

For detection of balancing selection, H_{01} and H_{02} are less conservative than H_{00}, although they are still conservative in the sense of either preserving the excess of older mutations, or reducing the number of younger mutations, or both, for feasible scenarios of human population history. The reason for this is that actual increase of the human population size was most likely larger than 10-fold growth over 5,000 generations. This makes H_{01} conservative, if the direction of departures from neutrality is towards excess of old mutations or reducing the number of young mutations or both (Fu 1996, Fu 1997).

Since H_{02} is always more conservative in the sense discussed above than H_{01}, then if H_{01} is conservative, so must be also H_{02}. H_{03} assumes the same demography as H_{02}, but takes recombination into account. It is therefore the most conservative and including a maximum number of genetic forces. Hence, departures from H_{03} should be interpreted as most likely caused by balancing selection. The results of testing for all loci, populations and null hypotheses are presented in Tables 8, 9, 10 and 11 for tests T, Z_{nS}, F^* and Q, respectively.

Outcomes of tests T and F^* against H_{00} are similar and significantly positive for ATM and RECQL. Such outcomes indicate that the polymorphism in loci considered exhibits an excess of old mutations, or a deficit of young mutations or both, compared to the neutral Wright-Fisher model (Fu 1997). At the same time, WRN and BLM do not show significant deviation from neutrality, although they deviate in the same direction as ATM and RECQL. Site by site comparison of human SNPs with corresponding ape sequences confirms the existence of old mutations in all loci.

For all helicases, a sample composed of 10 chromosomes from 2 chimpanzees, 1 bonobo and 2 gorillas indicates that human polymorphism is monomorphic among apes and we could treat the common ape haplotype as the ancestral sequence. For all genes considered such ancestral haplotype is present in human population at low frequencies.

Some of the mentioned above SNPs, like for example IVS15+33444t-c in RECQL, represent young mutations with mutated nucleotides present at very low frequencies, but other, such as IVS1-30638g-c or IVS19-30329g-t in the same gene, include derived mutations observed in the second, third and fourth most frequent haplotype. Such mutations, and especially those present in most common haplotypes, like IVS1-8213g-a in WRN or IVS1-20561t-c in BLM, are frequent and therefore likely to be old, consistent with the positive outcomes of Fu's *F*' (*r*, *r*') tests.

Table 4.3:8

Significance of the Tajima's *T* test for various null hypotheses. Dark, significant for 3-4 populations. Light, non significant for 1-2 populations. Unshaded, non significant for 3-4 populations

Gene	Population	Value	*T* (H_{00})	*T* (H_{01})	*T* (H_{02})	*T* (H_{03})
ATM	AfAm	2.42	*	***	*	*
	Caucasian	3.48	***	***	***	***
	Asian	2.55	*	***	**	**
	Hispanic	3.20	**	***	**	**
RECQL	AfAm	2.83	*	***	*	*
	Caucasian	3.10	**	***	**	**
	Asian	2.65	*	***	*	*
	Hispanic	2.93	**	***	**	**
WRN	AfAm	0.79	NS [a]	*	NS	NS
	Caucasian	1.26	NS	*	NS	NS
	Asian	1.36	NS	*	NS	NS
	Hispanic	1.10	NS	*	NS	NS
BLM	AfAm	2.06	NS	***	*	*
	Caucasian	2.50	*	***	**	**
	Asian	1.78	NS	**	NS	NS
	Hispanic	1.87	NS	**	NS	NS

***: $p < 0.001$, **: $0.01 > p \geq 0.001$, *: $0.05 > p \geq 0.01$, [a] NS (non significant): $p > 0.05$.

The excess of old mutations is also observed in ATM, and furthermore this locus contains a bi-allelic *trans*-polymorphism, shared by humans and chimpanzees at SNP IVS62+424g-a (shaded nucleotide in Table 7; note also framed nucleotides A in gorilla sequences, different from both chimp and human variations). If this between-species polymorphism is inherited from a common ancestor, the mutation must be several million years old (only balancing selection can preserve such old mutation) and even if it arose independently in humans and chimpanzees, the comparison of the most probable ancestral sequence, shared by chimp and

bonobo, with human haplotypes indicates old mutations having pattern similar to IVS1-30638g-c or IVS19-30329g-t in RECQL.

Table 4.3:9
Significance of the Kelly's Z_{nS} test for various null hypotheses. The meaning of shaded regions is the same as in Table 8

Gene	Population	Value	Z_{nS} (H_{00})	Z_{nS} (H_{01})	Z_{nS} (H_{02})	Z_{nS} (H_{03})
ATM	AfAm	0.29	NS [a]	*	NS	*
	Caucasian	0.47	*	**	**	**
	Asian	0.49	*	**	*	*
	Hispanic	0.45	*	**	*	*
RECQL	AfAm	0.24	NS	*	NS	NS
	Caucasian	0.36	NS	*	*	*
	Asian	0.52	*	**	*	*
	Hispanic	0.32	NS	*	NS	NS
WRN	AfAm	0.06	NS	?[b]	NS	NS
	Caucasian	0.10	NS	*	NS	NS
	Asian	0.18	NS	*	NS	NS
	Hispanic	0.12	NS	*	NS	NS
BLM	AfAm	0.12	NS	*	NS	NS
	Caucasian	0.18	NS	*	NS	NS
	Asian	0.17	NS	*	NS	NS
	Hispanic	0.15	NS	*	NS	NS

**: $0.01 > p \geq 0.001$, *: $0.05 > p \geq 0.01$, [a] NS (non significant): $p > 0.05$,
[b] ? (borderline): $p = 0.05$.

The phylogenetic tree (Fig. 3) reveals that the most ancient human hyplotypes 5 and 13 are very rare, and the most frequent haplotypes 2 and 31 with respective frequencies of 31% and 28%, belong to two separate clades. Having an indication about the excess of old mutations, it is possible to understand why the outcomes of T and F^* against H_{01} are significant for all loci. However, more interesting are the outcomes of testing against H_{02} and H_{03}, as they incorporate not only growth, but also substructure and, in the case of H_{03}, the effect of recombination. For both these hypotheses T and F^* are significant for ATM and RECQL, and F^* is also significant for BLM.

In Fig. 3, the first number indicates the reference number of haplotype and the second (if present) the frequency in percents (if absent the frequency is less than 1%). The number in parentheses gives the rank of the haplotype according to the global frequency in human population. For example the uppermost haplotype number 2 has frequency 31% and is the most frequent haplotype.

The pattern found in Kelly's Z_{nS} test outcomes (Table 9) is essentially the same as that in $F'(r,r')$ tests, yet the overall power seems to decrease. Still, the ATM and RECQL outcomes for the most reliable nulls are significant, although the significance is more evident in the

case of ATM. BLM and WRN are both non-significant. For ATM and RECQL loci Wall's *Q* outcomes (Table 11) are on the boundary of significance against H_{03} and non significant for WRN and BLM even against H_{01}.

Table 4.3:10
Significance of the Fu's *F** test for various null hypotheses. The meaning of shaded regions like in Table 8

Gene	Population	Value	*F** (H_{00})	*F** (H_{01})	*F** (H_{02})	*F** (H_{03})
ATM	AfAm	2.10	*	***	**	**
	Caucasian	2.60	**	***	***	***
	Asian	0.96	NS [a]	*	NS	NS
	Hispanic	2.47	*	***	**	**
RECQL	AfAm	1.68	NS	**	*	*
	Caucasian	2.30	**	***	**	**
	Asian	1.52	NS	**	*	*
	Hispanic	2.23	*	***	*	*
WRN	AfAm	0.21	NS	NS	NS	NS
	Caucasian	1.58	NS	*	*	*
	Asian	1.47	NS	*	*	NS
	Hispanic	0.05	NS	NS	NS	NS
BLM	AfAm	1.72	NS	***	*	*
	Caucasian	1.90	*	***	**	**
	Asian	1.58	NS	**	*	*
	Hispanic	1.65	NS	**	*	*

***: $p < 0.001$, **: $0.01 > p \geq 0.001$, *: $0.05 > p \geq 0.01$,

[a] NS (non significant): $p > 0.05$.

Nielsen (2001) suggests being conservative in conclusions about selection based on tests using only haplotype spectrum data, because other alternative hypotheses lead to similar results. The main alternative is that of population growth, which can be easily mistaken for a selection. These concerns, which are especially important in the case of selective sweeps as leading to an excess of young mutations (Fu 1997), are not directly applicable to this study, with samples displaying excess of old mutations. Furthermore the concerns of Nielsen are implicitly based on the assumption that testing is performed against H_{00}, *i.e.* classical Wright-Fisher model of neutral genetic drift in a panmictic constant-size population. In this study however, it was tested not only against H_{00}, but also against other null hypotheses formulated in a conservative way. If conservative rates of growth and migration have been chosen, then demographic factors should not obscure inferences.

Table 4.3:11

Significance of the Wall's *Q* test for various null hypotheses. The meaning of shaded regions is the same as in Table 8

Gene	Population	Value	Q (H_{00})	Q (H_{01})	Q (H_{02})	Q (H_{03})
ATM	AfAm	0	NS[a]	NS	NS	NS
	Caucasian	0.36	NS	*	*	*
	Asian	0.29	NS	*	NS	? [b]
	Hispanic	0.14	NS	?[b]	NS	NS
RECQL	AfAm	0.36	NS	*	?	?
	Caucasian	0.36	NS	*	?	?
	Asian	0.60	*	*	*	*
	Hispanic	0	NS	NS	NS	NS
WRN	AfAm	0	NS	NS	NS	NS
	Caucasian	0	NS	NS	NS	NS
	Asian	0	NS	NS	NS	NS
	Hispanic	0	NS	NS	NS	NS
BLM	AfAm	0	NS	NS	NS	NS
	Caucasian	0	NS	NS	NS	NS
	Asian	0	NS	NS	NS	NS
	Hispanic	0	NS	NS	NS	NS

*: $0.05 > p \geq 0.01$,

[a] NS (non significant): $p > 0.05$,

[b] ? (borderline): $p = 0.05$.

Population growth of 10-fold over 5,000 generations, assumed in H_{01}, is conservative, yet the conclusions based on testing against H_{01} alone could not be considered conservative, since H_{01} does not take into account the substructure of human population. Hypotheses H_{02} and H_{03} assumed, consistent with the out-of-Africa scenario, split of populations 5,000 generations ago and migration with normalized rate of $N_{end}\, m = 100$ between 4 demes in an island model with no isolation by distance. There can be some doubt whether such scenario is realistic for the data we used, coming from 4 subpopulations living in metropolitan area of Houston, TX, USA. First of all, these subpopulations genetically are not homogeneous themselves. Rather, they contain different levels of admixture.

Hispanic subpopulation is composed roughly of 60% European (Spanish) and 40% Native American genes, whereas African American subpopulation contains on the average about 75% African and 25% European genes (Chakraborty 1986). The Caucasian (European) and Asian population are less affected by the admixture and their loss of heterozygosity relative

to Hardy Weinberg expectations, as reflected in Wright's F_{ST}, is only about 1% (Budowle and Chakraborty 2001, Budowle *et al.* 2001).

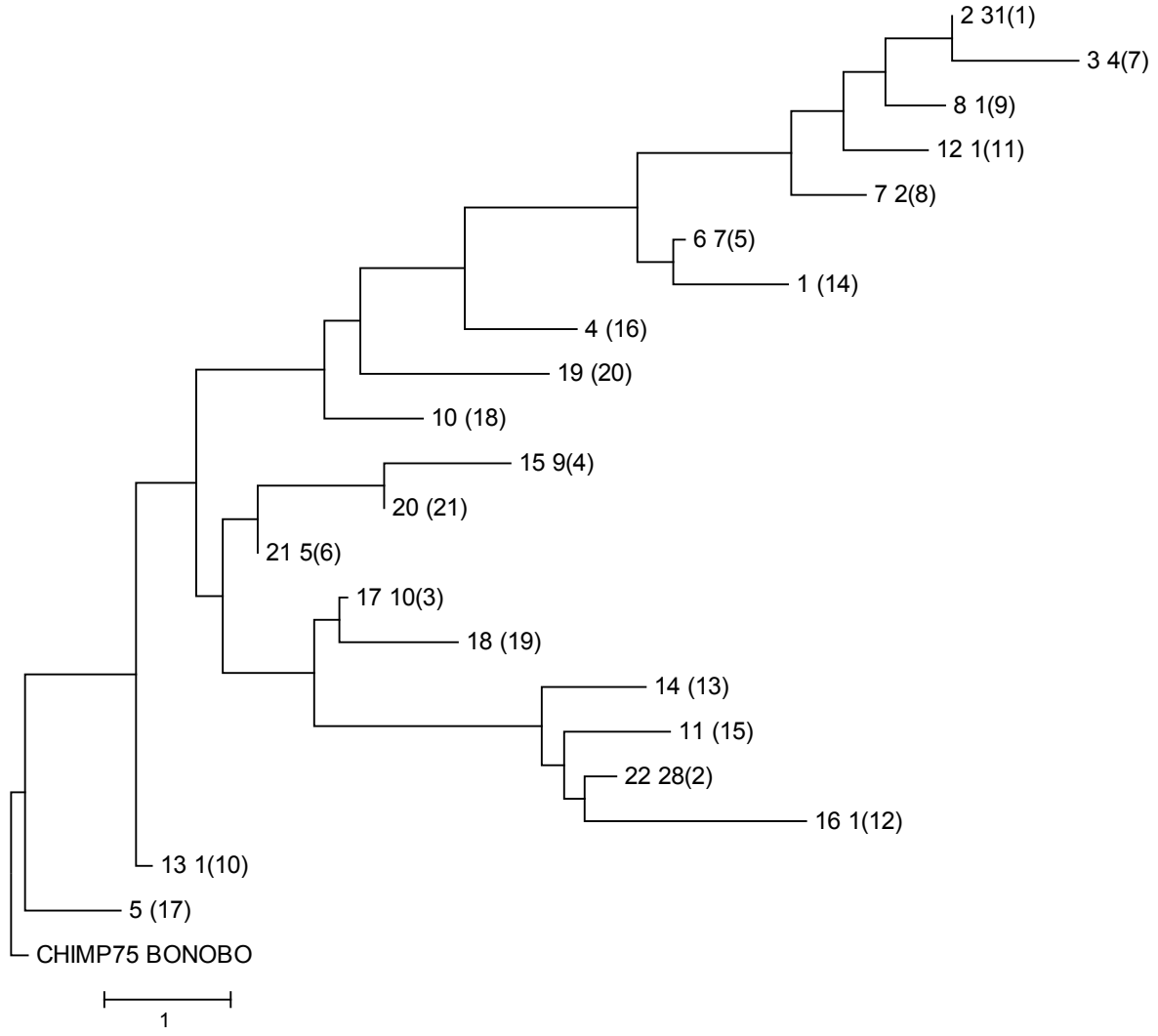


Fig. 4.3:3. The neighbor joining phylogenetic tree of the ATM haplotypes
Rys. 4.3:3. Drzewo filogenetyczne łączenia sąsiadów dla haplotypów ATM

Therefore, to check the sensitivity of obtained results with respect to the migration rate resulting in various gene admixtures, there were performed additional simulations for normalized migration coefficient ranging 100-fold, from 1 to 100. The corresponding change of critical values (results not shown) caused no dramatic difference in statistical significance for all populations and loci. In the hypotheses H_{02} and H_{03}, there were used the most conservative value of the parameter from the mentioned range. Interestingly, results showed that the strongest selection for all loci is found in Caucasians and, somewhat weaker, in Asians (both considered less admixed compared to African Americans and Hispanics).

Recombination rates for all loci are highest in African American population, consistent with the recent out-of-Africa scenario, assuming the largest effective population size of Africans.

It has been determined that the null hypothesis H_{03} is conservative and it incorporates alternatives other than selection. Therefore, statistically significant outcomes of practically all tests for ATM and RECQL loci (with tests T or F^* deviating in the direction of positive values) should be interpreted in the favor of overdominance selection. This type of selection preserves the polymorphism by rewarding heterozygotes. Hence, the question arises: What could be the molecular basis for selective scheme with rewarded heterozygotes? Which, selectively non neutral, functions and pathways are associated with these genes?

Together with BLM, ATM is one of the DNA repair proteins identified in a BASC (BRCA1-associated genome surveillance complex). Wang *et al.* (2000) suggested that BASC may serve as a sensor of abnormal DNA structures and as a regulator of the postreplication repair process. Cortez *et al.* (1999) showed that phosphorylation of BRCA1 by ATM may be critical for a proper response to DNA double-strand breaks and may provide a molecular explanation of the role of ATM in breast cancer. The interaction between ATM and BLM was confirmed by Beamish *et al.* (2002). By mutation analysis, they mapped the BLM-binding domain of ATM and ATM-binding domain of BLM.

Khanna *et al.* (1998) additionally found direct binding between ATM and p53 resulting in phosphorylation of serine 15 in p53, and thereby contributing to the activation and stabilization of p53 during the IR-induced DNA damage response. Lim *et al.* (1998) suggested that the large size of the protein and its multiple subcellular localization may indicate even more functions of the ATM. Recent studies (Yamaguchi *et al.* 2003) confirmed this hypothesis by association of ATM as a tumor suppressor in T-cell prolymphocytic leukemia.

As it was already stated, the function (other than helicase activity) of the second selectively non-neutral gene, RECQL, remains mainly unknown. Yet, there are some indications about its role when there is impaired BLM gene product activity (Wang *et al.* 2003). Homozygotic BLM deficient cells show slow-growth phenotype, a higher sensitivity to DNA-damaging agents and an approximately 10-fold increase in the frequency of sister chromatid exchange compared to wild-type cells. Analogous effect is not observed in cells with homozygotic RECQL knock-out. However cells with knock-out of both BLM and RECQL grow even more slowly than BLM(-/-) due to an increase of the proportion of dead cells in the population. The result suggests that RECQL is involved in cell viability if the BLM function is impaired (Wang *et al.* 2003). The cooperative role of RECQL is reflected also in its helicase activity. RECQL alone is able to unwind short DNA duplexes (less than 110 bp), but in the presence of human replication protein A (hRPA) as long as 500 bp substrates can be unwound.

There is no evidence as to which, if any, of these interactions could mechanistically explain the pattern of variation, which suggests overdominance at ATM and RECQL. However, recently Thomas and Kejariwal (2004) discovered that there existed a qualitative difference between the type of selection operating at loci involved in Mendelian diseases (like ataxia telangiectasia) compared to complex diseases. In the first type, the deleterious coding SNPs tend to occur at evolutionarily highly conserved amino acid positions, suggesting that they have a severe negative impact on the function of the protein (fold stability, active sites, etc.). However in genes implicated in complex diseases, including predisposition to malignancies, diabetes, etc.

Thomas and Kejariwal (2004) report possibility of greater (on average) positive selection pressure, since coding SNPs tend to occur at positions associated with the more subtle modulation of the protein function. In the light of the above finding, it is possible that in the case of pleiotropic genes, like ATM, involved in both Mendelian and complex diseases, the positive selection pressure is caused by evolutionarily advantageous heterozygotes required for overdominant selection. In such situation, overdominance may arise in a region of the gene with strong linkage disequilibrium and linked alleles *Ab* and *aB*, with *A* being a slightly positive allele, reducing predisposition to a complex disease, *a* and *B* being selectively neutral, for a complex and a Mendelian disease respectively, and *b* being strongly deleterious for homozygotes associated with a Mendelian recessive disease.

However, it remains unknown which specific modulations of gene functions can be causative for non-neutrality at these hypothetical alleles A and B. The results presented here, implicating presence of such alleles at ATM and RECQL loci, can encourage research with the goal of their identification and explanation of their impact on natural selection and evolution of these genes. In addition, these results support the hypothesis that the bi-allelic *trans*-specific polymorphism IVS62+424g-a discovered by group of Houstonian geneticists (Bonnen *et al.* 2000), shared by humans and chimpanzees at ATM locus, is the result of an ancient polymorphism present in a common ancestor of humans and chimps. This conclusion is based on the fact that all outcomes of our tests different from balancing selection would have rejected the hypothesis of common origin in favor of independent origins.

4.3.3. Artificial intelligence-based method

The required assumption for successful application of the AI-based methods is that a mosaic of test outcomes, making a direct inference so troublesome, contains enough information to differentiate between the existence of natural selection and its lack. The second prerequisite is the expert knowledge about presence of the selection for given combinations of neutrality test outcomes. Having those two, it is possible in principle to train

the knowledge retrieving system and, after successful testing, to use it for other genes for which the expert knowledge is unknown. The author has studied application of neural networks (in particular PNN) and three rough set approaches (CRSA, DRSA, and QDRSA) in the problem considered.

In experiment with application of PNN, in order to interpret the outcomes of the battery of mentioned seven tests, first there is applied complex multi-null-hypotheses methodology to obtained labels (balancing selection or no evidence of such selection) for given combination of tests results computed with the assumption of classical null hypothesis. The goal of the experiment was to prove that the information preserved in these test results (even computed without taking into account factors like population growth, recombination and population substructure) is valuable enough to obtain reliable inferences.

As a tool for this study probabilistic neural network was used. As presented in section 2.2.1, it is specialized radial basis function (RBF) network applicable almost exclusively for problems of classification in probabilistic uncertainty model. The network generates on its outputs likelihood functions $p(x|C_j)$ of input vectors x belonging to given class C_j. One should notice that likelihood functions also define random abstract classes defined in a probabilistic uncertainty model.

On the other hand, the likelihood function after multiplying it by prior probabilities of classes (approximated by frequencies of class representatives in a training set) and after dividing the result by the normalizing factor having the same value for all classes (and therefore negligible in a decision rule discriminating between classes) yields posterior probability $P(C_j|x)$ of the given class C_j, given the input vector x. However, this posterior probability is also the main criterion of a decision rule in a probabilistic uncertainty model implemented by Bayesian classifiers.

The mentioned above decision rule is very simple assuming the same cost of any incorrect decision (i.e. in the case considered treating equally false positive and false negative answers). It can be simply reduced to the choice of the class with maximum posterior probability $P(C_j|x)$.

Moreover, assuming the same frequencies of the representatives of all classes in a training set – what is the case in this study – the above rule is equivalent to the choice of the class with maximum likelihood $p(x,C_j)$. Since likelihood functions are generated by output neurons of probabilistic neural network, therefore to obtain a decision one has to drive inputs of the PNN with the given vector x, and choose the class corresponding to the neuron with the highest level of response.

The training of the probabilistic neural network is a one-epoch process, given the value of the parameter s denoting the width of the kernel in the pattern layer. Since the results of the classification are strongly dependent on the proper value of this parameter, in reality the one-

epoch training should be repeated many times in a framework used for optimization results with respect to *s*. Fortunately the shape of the optimized criterion in one dimensional space of parameter *s* in majority of cases is not too complex, with one global extreme having respectable basin of gravity. If the width parameter *s* is normalized by the dimensionality of the input data *N* in an argument of the kernel function, then the proper value of *s* is very often within a range from 10 to 10^{-1}. In this study, where there was applied the minimization of the decision error serving as a criterion, the optimal value of *s* proved to be 0.175.

Table 12 presents the results of PNN classification during jack knife cross validation for *s* equal to 0.175 (In the study three PNNs were trained, each with different width of the kernel function. In jack knife cross validation the PNN with $s = 0.175$ gave the best results). The decision error of this classifier in testing was equal only 6.25% with estimated standard deviation of this error equal to 0.067, proving very good classification abilities of the PNN.

Table 4.3:12
The results of jack-knife cross validation procedure for the probabilistic neural network with parameter $s = 0.175$ (93.5% correct decisions)

Test Number	Number of correct decisions	Percentage of correct decisions	Decision error
1	2	100%	0
2	1	50%	0.5
3	1	100%	0
4	2	100%	0
5	2	100%	0
6	2	100%	0
7	2	100%	0
8	2	100%	0
Average	15/16	93.75%	0.0625

To compare three rough set-based approaches (CRSA, DRSA, and QDRSA) applied for testing of balancing selection in four genes involved in human familial cancer, consider the information system $S = (U, Q, V_q, f)$ in which $Q = C \cup \{d\}$. The haplotypes for particular loci were inferred and their frequencies were estimated by using the Expectation-Maximization algorithm (Polańska 2003). The results of tests *T*, *D**, *F**, *S*, *Q*, *B* and Z_{nS}, together with the decision concerning the evidence of balancing selection based on multi-null methodology, are given in a Table 13.

The rough set based analysis of the Decision Table 1, reveals that there exist two relative reducts: RED_1={*D**, *T*, Z_{nS}} and RED_2= {*D**, *T*, *F**}. It is clearly visible, that the core set is composed of tests *D** and *T*, whereas tests Z_{nS} and *F** can be chosen arbitrarily, according to the automatic data analysis. However, since it is known, that both Fu's tests *F** and *D** are the examples of tests belonging to the same family, and therefore their outcomes are rather strongly correlated, it is advantageous to choose Kelly's Z_{nS} instead of *F** test. It is so, because Z_{nS} outcomes are theoretically less correlated with outcomes of test *D**, belonging, as

it was stated above, to the core and therefore required in any reduct. The Decision Table 1 with reduced set of conditional attributes to the set RED_1 is presented in Table 14.

Table 4.3:13

Decision Table 1. The outcomes of the statistical tests for the classical null hypothesis

		D^*	B	Q	T	S	Z_{nS}	F^*	Balancing Selection
ATM	AfAm	*	NS	NS	*	NS	NS	*	Yes
	Cauc	*	NS	NS	**	**	*	**	Yes
	Asian	NS	NS	NS	*	NS	*	NS	Yes
	Hisp	*	NS	NS	**	NS	*	*	Yes
RECQL	AfAm	NS	NS	NS	**	NS	NS	NS	Yes
	Cauc	*	NS	NS	**	NS	NS	**	Yes
	Asian	NS	*	*	*	NS	*	NS	Yes
	Hisp	*	NS	NS	**	NS	NS	*	Yes
WRN	AfAm	NS	NS	NS	NS	NS	NS	NS	No
	Cauc	*	NS	NS	NS	NS	NS	NS	No
	Asian	*	NS	NS	NS	NS	NS	NS	No
	Hisp	NS	NS	NS	NS	NS	NS	NS	No
BLM	AfAm	NS	NS	NS	NS	NS	NS	NS	No
	Cauc	NS	NS	NS	*	NS	NS	*	No
	Asian	NS	NS	NS	NS	NS	NS	NS	No
	Hisp	NS	NS	NS	NS	NS	NS	NS	No

Table 4.3:14

Decision Table 2, in which the set of tests is reduced to relative reduct RED_1 composed of tests: D^*, T, and Z_{nS}

		D^*	T	Z_{nS}	Balancing Selection
ATM	AfAm	*	*	NS	Yes
	Cauc	*	**	*	Yes
	Asian	NS	*	*	Yes
	Hisp	*	**	*	Yes
RECQL	AfAm	NS	**	NS	Yes
	Cauc	*	**	NS	Yes
	Asian	NS	*	*	Yes
	Hisp	*	**	NS	Yes
WRN	AfAm	NS	NS	NS	No
	Cauc	*	NS	NS	No
	Asian	*	NS	NS	No
	Hisp	NS	NS	NS	No
BLM	AfAm	NS	NS	NS	No
	Cauc	NS	*	NS	No
	Asian	NS	NS	NS	No
	Hisp	NS	NS	NS	No

After a reduction of the set of informative tests to a set $RED_1=\{D^*, T, Z_{nS}\}$, we considered the problem of a coverage of the discrete space generated by these statistics, by the examples included in the training set. The results are given in a Table 15, in which the domain of each of the test outcome (coordinate) is composed of three values: ** (strong statistical significance $p < 0.01$), * (statistical significance $0.01< p < 0.05$), and *NS* (non significance $p > 0.05$). The given point in a space is assigned to: *S* (the evidence of balancing selection), *N* (no evidence of balancing selection) or empty cell (point not covered by the

training data). The assignment is done based on raw training data with conditional part reduced to the relative reduct RED_1 . Note, that the percentage of points, covered by training examples, is only 30%.

Table 4.3:15
The discrete space of three tests: D^*, T, and Z_{nS}, based on Decision Table 2

		T								
		**			*			*NS*		
		Z_{nS}			Z_{nS}			Z_{nS}		
		**	*	*NS*	**	*	*NS*	**	*	*NS*
D^*	**									
	*		*S*	*S*			*S*			*N*
	NS			*S*		*S*	*N*			*N*

The next step was to apply the notion of the relative value reducts to particular decision rules in the Decision Table 2. The resulting Decision Table 3 is presented in a Table 16.

Table 4.3:16
Decision Table 3, based on relative value reducts for three tests: D^*, T, and Z_{nS}

		D^*	T	Z_{nS}	Balancing Selection
ATM	AfAm	*	*		Yes
	Cauc		**		Yes
	Asian			*	Yes
	Hisp		**		Yes
RECQL	AfAm		**		Yes
	Cauc		**		Yes
	Asian			*	Yes
	Hisp		**		Yes
WRN	AfAm		NS		No
	Cauc		NS		No
	Asian		NS		No
	Hisp		NS		No
BLM	AfAm		NS		No
	Cauc	NS	*	NS	No
	Asian		NS		No
	Hisp		NS		No

Table 17 presents information analogous to Table 15, however the coverage of points is based on the number of points which are classified with the use Decision Table 3. One should notice that the percentage of covered by algorithm points is 74%, however since 11% (denoted with "-") is classified as both with and without the evidence of balancing selection, therefore only 63% of the points could be treated as really covered.

Based on this Decision Table 3, the $Algorithm_{CRSA}$ has been obtained using CRSA.. Note that this algorithm is simplified as compared to the algorithm which corresponds to the Decision Table 2. At the same time, it is more general, what can be observed in a Table 17, as compared to Table 15. In the algorithm, the outcomes of neutrality tests are designated as *NS*, *S*, and *SS* for non-significant, significant, and strongly significant, respectively.

Table 4.3:17
The discrete space of three tests: D^*, T, and Z_{nS}, based on Decision Table 3

		T = **			T = *			T = NS		
		Z_{nS} = **	Z_{nS} = *	Z_{nS} = NS	Z_{nS} = **	Z_{nS} = *	Z_{nS} = NS	Z_{nS} = **	Z_{nS} = *	Z_{nS} = NS
D^*	**	S	S	S					-	N
	*	S	S	S	S	S	S		-	N
	NS	S	S	S		S	N		-	N

Algorithm$_{CRSA}$, (Cyran 2009d)

```
BAL_SEL_DETECTED   = False
BAL_SEL_UNDETECTED = False
CONTRADICTION      = False
NO_DECISION         = False

if T = SS or (T = S and D* = S) or ZnS = S then
   BAL_SEL_DETECTED = True
if T = NS or (T = S and D* = NS and ZnS = NS) then
   BAL_SEL_UNDETECTED = True

if BAL_SEL_DETECTED and
   BAL_SEL_UNDETECTED) then
   CONTRADICTION = True
if not(BAL_SEL_DETECTED) and
   not(BAL_SEL_UNDETECTED) or
   CONTRADICTION then
   NO_DECISION = True
```

The algorithm generated by DRSA, called *Algorithm*$_{DRSA}$ is as follows

Algorithm$_{DRSA}$, (Cyran 2009d)

```
at_least.BAL_SEL_DETECTED  = False
at_most.BAL_SEL_UNDETECTED = False
CONTRADICTION              = False
NO_DECISION            = False

if T >= SS or (T >= S and D* >= S) or ZnS >= S then
   at_least.BAL_SEL_DETECTED = True
if T <= NS or (T <= S and D* <= NS and ZnS <= NS) then
   at_most.BAL_SEL_UNDETECTED = True

if at_least.BAL_SEL_DETECTED and
   at_most.BAL_SEL_UNDETECTED then
   CONTRADICTION = True
if not(at_least.BAL_SEL_DETECTED)
   and not(at_most.BAL_SEL_UNDETECTED) or
   CONTRADICTION then
   NO_DECISION = True
```

It happened that the algorithm generated by QDRSA *Algorithm*$_{QDRSA}$ is identical to *Algorithm*$_{DRSA}$ when the whole universe U of the information system S is used for generation of the algorithm. However, if the universe of the information system S is divided into two sets of rules, those used for information retrieval in the process of generating the decision algorithm, and those left for testing, then the resulting algorithms generated by DRSA and

QDRSA are different in some cases. Below only these algorithms which differ between the two approaches are presented.

If the information about RECQL gene is excluded from the information system *S* and it is left for testing in crossvalidation process, then the DRSA and QDRSA generate the algorithms $Algorithm_{DRSA}$(-*RECQL*) and $Algorithm_{QDRSA}$(-*RECQL*), respectively. Since the general structure of both algorithms is identical to that of $Algorithm_{DRSA}$, only two crucial if-then rules (the ones after four initialization assignments, and before two contradiction/no-decision determining if-then rules) are presented below.

$Algorithm_{DRSA}$(-*RECQL*)

```
...
if (T >= S and D* >= S) or Zns >= S then
   at_least.BAL_SEL_DETECTED = True
if T <= NS or (D* <= NS and Zns <= NS) then
   at_most.BAL_SEL_UNDETECTED = True
...
```

$Algorithm_{QDRSA}$(-*RECQL*)

```
...
if {T >= SS} or
   (T >= S and D* >= S) or Zns >= S  then
   at_least.BAL_SEL_DETECTED = True
if T <= NS or (D* <= NS and Zns <= NS) then
   at_most.BAL_SEL_UNDETECTED = True
...
```

It is visible that the difference is the existence of one more condition in the rule describing the detection of balancing selection. This condition reads "if the outcome of Tajima test is at least strongly statistically significant". It occurs in $Algorithm_{QDRSA}$(-*RECQL*), because the condition $T = SS$ is the result of application of the relative value reduct for one of the rules in the information system *S*(-*RECQL*) analyzed with QDRSA indiscernibility relation (2.3:5). After changing the condition in QDRSA to $T >= SS$, this condition is still not dominated by any other conditions detecting balancing selection. Since it is not dominated it must remain in the final decision algorithm presented above.

However, this is not the case in DRSA. This latter approach, when considering the dominance of the decision rules for the class *at-least.BAL-SEL*, compares the original (i.e. not reduced with the relative value reduct notion) condition (A) $D^* >= S$ and $T >= SS$ and $Z_{nS} >= S$ with another original condition (B) $D^* >= S$ and $T >= S$ and Z_{nS} $Z_{nS} >= NS$, instead of comparing (like QDRSA does) the condition (a) $T >= SS$ with condition (b) $D^* >= S$ and $T >= S$, being the results of application of the relative value reducts in QDRSA-sense to the original conditions (A) and (B), respectively.

It is clear, that the rule with the condition (A) is dominated by the rule with the condition (B), and therefore the condition (A) seemed to be redundant in DRSA-sense for the class *at-*

least.BAL-SEL. However, the rule with the condition (a) is not dominated by the rule with the condition (b) and this is the reason why condition (a) is present in the $Algorithm_{QDRSA}$(*-RECQL*), while it is absent in $Algorithm_{DRSA}$(*-RECQL*). The conditions (B) and (b) in both approaches are necessary and they are reduced to the condition (b) present in both algorithms.

Finally, consider what is the influence of inclusion of the condition $T >= SS$ to the $Algorithm_{QDRSA}$(*-RECQL*). When this algorithm is applied for the interpretation of neutrality tests for RECQL gene, i.e. the gene which was not present in the information system *S*(*-RECQL*), the decision error is reduced from 0.25 to 0 for four populations. When the full jack-knife method of the crossvalidation is applied, then the decision error is reduced from 0.313 with DRSA, what seems rather unacceptable, to 0.125 with QDRSA. It is important to mention that at the same time QDRSA *NO-DECISION* results have increased from 0 to 0.188, however in the case of screening procedure for which this methodology is intended, the unsure decision is also an indication for the more detailed study with the use of multi-null hypotheses methodology.

4.4. Conclusions

Population geneticists have developed quite a number of statistical neutrality tests which serve to deny at given significance level the Kimura's model of neutral evolution described in section 4.1. Hence, in the post-genomic area researchers are armed with quite a number of statistical tests (see section 4.2) whose purpose is to detect signatures of natural selection operating at the molecular level. Positive signals generated by these tests, given in detail in section 4.2, can be interpreted as caused be the presence of natural selection. In the case study considered in section 4.3 there have been used the following neutrality tests: Tajima's *T*, Fu's *D** and F_s, Wall's *Q* and *B*, Kelly's Z_{nS} and Strobeck's *S*.

However, because of such factors like recombination, population growth, and/or population subdivision, the appropriate interpretation of the test results is very often troublesome (Nielsen 2001). When the given gene is tested with the use of aforementioned tests, some of them can give positive, while others generate negative signals. Moreover, positive signals can be caused by population expansion or geographical structure of the population. On the other hand the signatures of actual natural selection can be suppressed by the recombination. All these factors make the proper interpretation hard, and not necessarily univocal. The problem is that mentioned departures from selectively-neutral classical model

(i.e. model with panmictic, constant in size population with no recombination) can produce similar results for some of these tests to results produced by the existence of natural selection.

Nevertheless, since the time of Kimura's famous book (Kimura 1985) until present, geneticists are searching for signatures of natural selection, treating proposed by Kimura model of neutral evolution at molecular level as a convenient null hypothesis, which is not fulfilled for particular loci under detectable selection. By moving the emphasis form selective forces to random genetic drift and neutral mutations, the neutral theory of molecular evolution gave birth to mentioned neutrality tests, which treat this theory as a null model, and statistically significant departures from it, discovered in loci under study, can be interpreted in a favor of natural selection. The existence of a rare, positive selection has been confirmed for example in a ASPM locus that contributes to the size of brain in primates (Evans et al. 2004, Zhang 2003).

An interesting example of another type of selection, called balancing selection (see section 3.4), has been detected by the author in ATM and RECQL loci (see section 4.3). To overcome serious interpretation difficulties while searching for the selection in ATM, RECQL, WRN and BLM, i.e. in four human familial cancer genes, the author has proposed an idea of so called multi-null-hypotheses methodology (part of this methodology was published in Cyran et al. 2004). However, this methodology is not appropriate for fast detection because of long lasting computer simulations required for estimating critical values under non-classical null hypotheses.

Yet, armed with reliable conclusions about balancing selection at ATM and RECQL and no evidence of such a selection at WRN and BLM, after time consuming search with the use of computer simulations, the author has proposed the usage of machine learning methodology, based only on knowledge of critical values for classical null hypotheses (see section 4.3.3). Fortunately, critical values for classical nulls are known for all proposed non-neutrality tests, and therefore outcomes of such tests can be used as inputs for artificial intelligence classifiers without additional computer stochastic simulations of alternative models.

In this methodology, described in section 4.3.3, the battery of tests outcomes is considered as a set of conditional attributes and the expert knowledge is delivered by application of the multi-null hypotheses method for some small amount of genes. After crossvalidation of the model, the decision concerning other genes can be done based on testing only against classical null hypotheses and application of the decision algorithm inferred with AI-based methodology. Such strategy does no need intensive computer

simulation, and therefore is much more time-efficient as compared to multi-null hypotheses approach.

The results of application of rough set based theory for knowledge acquisition and processing were published in (Cyran 2007a) for CRSA, (Cyran 2010) for CRSA and DRSA, and (Cyran 2009d) for QDRSA. In Cyran (2009b) the author presented results of another study, based on the application of probabilistic neural network (PNN) for the detection of natural selection at molecular level. The advantage of the last proposed methods is that it not so time consuming and due to good recognition abilities of probabilistic neural networks it gives low decision error levels in cross validation (see section 4.3.3 for results).

The comparison of CRSA with DRSA for this particular purpose is described in section 4.3.3, where it is proved that neither CRSA nor DRSA generates decision algorithm which is optimal for the problem considered. The proof is done by a simple demonstration of another algorithm which is Pareto-preferred over both mentioned approaches. This algorithm can be obtained with QDRSA, the novel method proposed by Cyran (2009d).

The comparison of QDRSA with CRSA gives the favor to the first when the preference-order is present in conditional and decision attributes. The resulting decision algorithms in QDRSA are more general, i.e. they cover more points of the input space. Moreover, in many cases, because of possible domination of some QDRSA conditions over some other ones, the decision algorithms are shorter as compared to CRSA. However, because the domination is checked after the application of relative value reducts, the negative effect (characteristic to DRSA) of omitting the important condition from the decision algorithm (as it was shown in section 4.3.3 in the illustrative example concerning the search for signatures of natural selection operating at molecular level) is not present in QDRSA.

5. HUMAN EVOLUTION

5.1. Foundations

In the last decade a lot of relevant discoveries has been made in the area of origin of our species. These discoveries vary from fossils dated to several million years old, like skeleton of the *Pierolapithecus catalaunicus* being the early Great Ape from middle Miocene (Moya-Sola et al. 2004) or a few million years younger skeletons of *Sahelanthropus tchadensis*, *Orrorin tugenensis*, *Ardipithekus ramidus* and *Australopithecus anamensis* claimed to be our extinct antecessors living in Pliocene (Leakey and Walker 2003, Tattersall 2003a), to fossils as young as several thousand years old LB1 skeleton of *Homo floresiensis* (Brown et al. 2004).

The latter is especially intriguing, as it is representative of the order *Homo* probably different from our own species and being alive in Indonesian island in late Pleistocene, only about 38,000-18,000 years ago (Morwood et al. 2004) *i.e.* after *Homo sapiens* appeared in the region (55,000-35,000 years ago). Due to the height of the body (approximately 1m), and because of the size of the brain (about 380cm^3) *H. floresiensis* exhibits the most extreme case of the genus *Homo* and hardly matches any of two main interpretations of human origins.

The first interpretation, called the multiregional hypothesis (Walpoff 1999), assumes that modern humans evolved from the *H. erectus* species, which dispersed over the Old World more than one million years ago. In this hypothesis the genetic flow between these archaic human populations was so strong that it is justified to talk about one large-scale evolutionary process, which led from *H. erectus* to *H. sapiens*. The competing theory, known as the recent out-of-Africa origin hypothesis (Wilson and Cann 1992), assumes that there was very limited gene flow between archaic human populations which emerged from *H. erectus* and a population of anatomically modern humans, which left Africa about 100,000 years ago and spread through the Old World in subsequent tens thousand of years, reaching the New World through the Bering Sea frozen in Ice Ages, some 20,000 years ago. The debate between these

two models is still open, although the recent out-of-Africa hypothesis is considered by majority as the one, which better reflects the genetic record of humans (Jobling et al. 2004).

While it requires some time (and perhaps new discoveries) to give the coherent explanation of the *H. floresiensis* within (slightly?) rewritten human origin hypotheses, the early conclusions of evident isolation of small-bodied humans, seem to contradict the multiregionality. Mirazon Lahr and Foley (2004) express this fact even stronger writing in Nature that "*H. floresiensis* puts yet another (the last?) nail in the multiregional coffin". There are some doubts whether multiregionalists become convinced. They claim they posses strong paleoanthropological support for multiregional evolution of humans in continuity of anatomical features (especially in Asia, but also in Australia and Europe) before and after arrival of modern humans dispersing out of Africa (Thorne and Wolpoff 1992). Indeed, assuming the lack of interbreeding between archaic (autochthons) and modern (invaders) humans, it is hardly to explain the fact that some bone features of Australians, being clearly distinctive from Africans, are present in Australian fossils before and after the appearance of modern humans in the region.

Tattersall (2003b) does not agree with this interpretation and considers *Homo erectus* as the local evolutionary dead path, and Wilson and Cann (1992) address the problem indicating that mentioned bone features are not necessarily independent and selectively neutral. They suggest that successive re-evolution of similar bone patterns is plausible in similar environmental conditions. Still, the relatively short time required for replication of changes, makes this explanation at least disputable, especially having in mind that also some nuclear genes support different histories as compared to those inferred from mitochondrial DNA (mtDNA) (Hey 1997).

Nevertheless, due to the ease of PCR amplification of mtDNA present in a one cell in multiple copies, mtDNA-based inferences are an important source of our knowledge about origin of modern humans. This is true even more in the light of conflicting inferences yielded based on multiple autosomal microsatellite loci. Kimmel et al. (1998) suggested that extensive population growth has occurred in Asia and Europe and not in Africa, whereas Reich and Goldstein (1998) inferred just opposite. Therefore, successful sequencing of the mtDNA (yielding more unique results due to the lack of recombination) from Neanderthal fossils became the mile stone in revealing our evolutionary paths.

For example, until recently, the estimation of the mitochondrial mutation rate could rely only on human-chimpanzee divergence data. However due to relatively long time to this divergence, all estimates of this time were very inaccurate ranging from 4 to 9 million years (O'Connell 1995) – with the most probable value of 6 million years. Consequently estimated mutation rate could not be accurate and so is true with mitochondrial Eve (mtEve) epoch. O'Connell (1995) proved that the same genetic diversity of modern humans applied to his

branching process based model can give estimates of the mtEve epoch between 700 thousand even up to 1.5 million years.

These results were very different from those obtained with the use of phylogenetic trees estimated to 280 and 200 thousand of years by Hasegawa and Horai (1991) and Wilson and Cann (1992) respectively. The difference was not only due to very small sample size used by O'Connell (just 19 individuals resulting in too large genetic diversity of contemporary humans as compared to more recent data) but mainly due to insufficient concordance of his model with actual evolution of humans for times of order of million of years. In his paper O'Connell indicated also decreasing reliability of outgroup based methods when the outgroup is not close enough in genetic distance to the considered sample. Summarizing, until recently, mtEve dating estimates were dependent on inaccurate inference about human-chimpanzee divergence time and furthermore, they depended to great extent on the method applied for inferring.

When in 1997 (Krings *et al.* 1997) for the first time the mtDNA was sequenced from *Homo neanderthalensis* dated to be alive about 40,000 years ago (Schmitz *et al.* 2002), only less than 400 base pairs were sequenced. The next successful sequencings of Neanderthal mtDNA in 1999 (Krings *et al.* 1999) 2000 (Ovchinnikov *et al.* 2000, Krings *et al.* 2000) confirmed the accuracy of the first experiment. Since then, the mtDNA divergence rate no longer has to be guessed relying on the assumption of its constancy over a few million years, and problematic dating of human-chimpanzee split.

In 2004 the four additional Neanderthal fossils yielded mtDNA sequences together with five early modern humans fossils (Serre et al. 2004) and the results were in full concordance with previous sequencing efforts. What is also important, fossils sequenced by Serre et al. (2004) contained examples (Vandija 77, Vandija 80, Mladeč 25c, Mladeč 2) considered by multiregionalists as "transitional" between Neanderthals and early modern humans due to some morphological features (Smith 1984, Frayer 1986, 1992, and Wolpoff 1999). Yet the mtDNA proved to be of Neanderthal type for Vandija fossils considered as Neanderthals, and of modern human type for Mladeč fossils, considered as modern humans. This is exactly, what is expected by recent out-of-Africa model, suggesting that some morphological features shared by mentioned fossils can be results of similar environmental influence or could arise just by chance without strong genetic flow between Neanderthals and early modern humans.

Serre et al (2004), apart from reporting these results try to estimate the upper limit of possible Neanderthal admixture to early modern humans, consistent with mtDNA testimony. They use a coalescence method in three different demographies: (i) constant population size and population growth (ii) before and (iii) after potential point of Neanderthal admixture respectively. The numerical value of the estimate equal to 25 percent is given only for the simplest case of population constant size, known however to be unrealistic. In section 5.4,

similar (but indicating smaller admixture) limit it estimated, using branching process methodology. Interestingly branching processes have been recently also used for inferring the age of the primate last common ancestor based on archeological stratification and the number of species known to live in a given period (Tavare et al. 2002).

The results obtained by the author (section 5.4, see also Cyran and Kimmel 2005, Cyran 2010) further reduce the hypothetical Neanderthal mtDNA admixture to early modern humans gene pool. Even better estimates are possible when the history of human population inferred from archeological studies correlating Aurignacian, Chatelperronian and Gravettian cultures with Neanderthals or modern humans (Mellars 2004), as well as the influence of the Ice Ages on demography (Forster 2004) will yield more reliable estimates of the population size in different regions of the globe and corresponding time-inhomogeneous branching processes will be used.

As it was stated, human evolution at molecular level is reflected in the genome record. However, it is often hard to interpret this record, because a population under consideration could undergo periods of expansions, which, if undetected, could lead to erroneous inferences. Therefore, the problem of detecting past population growths become one of crucial issues in contemporary population genetics This problem is addressed in section 5.2 using the microsatellite markers.

Microsatellites are short tandem repeats, STRs (Renwick et al. 2001, Agrafioti and Stumpf 2007, Vowles and Amos 2006), which are quite abundant in genomes and undergo relatively fast mutations. Therefore, they are suitable for testing the evolution of populations rather than emergence of species, and no doubt, they have found applications in various tests for population detection. Using such data the author has proposed a new statistical test, which has greater power for detection of population growth than other available microsatellite based methods (see section 5.2 for details).

Moreover, some genes were under strong pressure of natural selection (the efforts aiming to search the signatures of such selection have been described in section 4.3), while genetic variation in others is mainly the result of the genetic drift (see section 3.2) and the selectively neutral mutations (see section 3.3 and 4.1). If the gene under consideration is exhibiting signatures of natural selection (see section 3.4) then some variants of it must be more or less fit to the environment. Very often it is associated with some disorder having genetic background, but in some cases it is responsible for the development of the species.

The best known example of the latter is the ASPM gene responsible for the brain size in primates, including humans (Zhang 2003). As presented in section 3.4, and also in sections 4.2 and 4.3, there is also balancing selection in which the heterozygotes (i.e. organisms having different alleles at two homologues chromosomes) are more fit than any homozygotes (i.e. organisms having identical variants at both homologues chromosomes). This is the case

with human sickle cell anemia which is caused by two identical copies of mutated allele. However, if this allele is present in heterozygote together with wild-type allele, then the carrier of one copy of mutant allele, not only does not suffer sickle cell anemia, but also this individual is able to generate successful immune response to the malaria. Therefore, on malaria endemic regions the mutant allele is frequent, despite it is responsible for severe disorder in homozygotes.

The indices of genetic variation, including allele distribution, heterozygosity or linkage disequilibrium, are affected by the population history. Therefore a lot of effort has been spent by statistical geneticists to estimate the long-term demographic history of populations belonging to various species. For this purpose many statistical tests detecting past population expansion have been proposed, for example King *et al*. (2000), Bjorklund (2003), Laan *et al*. (2005), Cyran and Myszor (2008b). Section 5.2 details the efforts in this field, and in particular, presents original neural network-based test (Cyran and Myszor 2008b, 2008c) with power exceeding powers of other known tests for detecting past population expansion.

In particular, the interest in our own history induced in the last decades the research focused on inferring the human population history (Polański and Kimmel 2003). DNA sequences which reflect genetic diversity taken from many qualitatively different loci of *H. sapiens* and *H. Neanderthalensis* have been analyzed. These analyses include for example studies of maternally inherited mitochondrial DNA (mtDNA) (Serre *et al*. 2004, Krings *et al*. 2000, Krings *et al*. 1999, Krings *et al*. 1997, Rogers 1995), paternally inherited Y chromosomes (Jobling 2001, Thompson *et al*. 2000), X chromosomes (Wooding and Rogers 2000), autosomal DNA sequences (Yu *et al*. 2001, Noonan *et al*. 2006, Pennisi 2007), nuclear short tandem repeats (STRs) (Kimmel *et al*. 1998), or protein sequences including β-globin (Harding *et al*. 1997, Fullerton *et al*. 1994), pyruvate dehydrogenase alpha 1 (PDHA1) (Hey 1997) or Duchenne muscular dystrophy gene product (DMD) (Zietkiewicz *et al*. 1998).

Despite these and similar efforts the problem of human population trajectory is still open and thus there is a growing interest in studies on how sensitive are genetic variation indices to departures from assumed in different models population histories. Moreover, applicability of methods for calculating the distributions of the time to coalescence is limited to the model within which they have been formulated.

The most widely used models assume simplifications such as multinomial sampling or deterministic population size. The question arises how robust they are for populations evolving stochastically. One interesting example which comprises stochasticity is O'Connell limit theory of genealogy in branching processes. This problem is explored in section 5.3. In particular, it is considered there how fast, in terms of number of generations, the limiting distributions of O'Connell are adequate descriptions of transient distributions.

To answer the problem extensive simulations of slightly supercritical branching processes were performed and the results are compared with O'Connell limits. Furthermore, coalescent computations under the Wright-Fisher model are compared with limiting O'Connell results and with full genealogy-based expectations. These expectations are used to estimate the age of the root of mitochondrial polymorphism of modern humans (or in other words to date the Mitochondrial Eve epoch), based on mtDNA sequenced from living humans and Neanderthal fossils.

Finally the problem of Neanderthal admixture in a gene pool of Upper Paleolithic anatomically modern humans is considered in section 5.4. The methodology applied accounts for the effect of the genetic drift, which could eliminate the hypothetical Neanderthal mtDNA admixture until present. To model the demography, the slightly supercritical Markov's BP based on the O'Connell model has been proposed. Relying on relatively fast convergence to the O'Connell's limiting properties it was possible to estimate the time of extinction of the Neanderthals relatively to the time of the root of the mtDNA polymorphism of modern humans.

The results of the study presented in section 5.4 indicate that the maximum hypothetical contribution of Neanderthal mtDNA which could be eliminated by the genetic drift at 0.05 significance level is about 12%. Moreover, the expected value of the admixture has been estimated to be about 4%. Relevance of the research considered in section 5.4 lies in treating mtDNA-based studies as complementary approaches to those based on nuclear DNA sequenced by the Neanderthal genome project.

5.2. Inferring demography

Coalescent theory (see section 3.5) enables creating huge amounts of samples in quite a short time (Marjoram and Wall 2006), yet its methods were developed some years ago when computers were rather expensive and possessed relatively low computational power. Over the last years the situation has changed due to invention of multi-core processors and overall progress in technology, which makes contemporary hardware highly efficient in computations and available at reasonable price. What is more, some recent research shows that given circumstances, coalescent methods might return different results than time-forward simulation approach.

In both coalescent-based and time-forward simulation methods it is often desired to obtain sample from population with experienced changes in amount of individuals between generations. One interesting application is to simulate changes of chosen genetic markers caused by mutation process. In the case of genetic markers the microsatellites can be used.

These are short strains of DNA build from repeating motifs of length 2-6 nucleotides (Renwick et al. 2001). Length of microsatellite is denoted by an amount of such repeated motifs, usually 60 or so (Goldstein and Pollock 1997).

Common mutation in microsatellites are changes in the amount of repeated motifs, i.e. change in the length of a microsatellite (Sia et al. 2000). Usually there is used one-step symmetric stepwise mutation model (SSMM), in which microsatellite might change length by one, with additional assumption that the probability of addition and deletion of one repeating motif is equal (Kimura and Ohta 1978). Microsatellites became popular because of their relative high mutation rate (about 10^{-4} – 10^{-5}), and the fact that they are spread all over genome (Zhivotovsky et al. 1997) – in human genome more than 10 000 microsatellites have been identified (Agrafioti and Stumpf 2007). Additionally, most of them is in non coding DNA, so according to neutral model of molecular evolution, they probably do not have influence on reproductive capabilities of individuals. Furthermore, microsatellites are easy in mathematical analysis.

During the research work performed by Cyran and Myszor (2008a), there was created a series of populations that underwent different kind and magnitude of growth. To simulate development of the population the model providing dynamic description of the evolution was formulated. It was based on the Wright – Fisher model (see section 3.2), which, in the most often used version, assumes (Hein et al. 2005):

- discrete and non overlapping generations,
- haploid individuals in populations,
- constancy of population size,
- equilibrium fitness of individuals in the population,
- lack of geographical or social structure in the population,
- no recombination in the population.

Because there were simulated populations whose size was changing in time, the applied W-F model allowed for changes in population size. The experiments concerned the Y (Bachtrog and Charlesworth 2001) chromosome or mtDNA (Eyre-Walker and Awadalla 2001) in order to eliminate the recombination issues and provide haploid individuals. When new generation was created the old one was deleted so there were no overlapping generations. During creation of new individual all parents could be chosen with equal probability, what eliminated problems of individuals' fitness and geographical or social structure.

In time forward simulation, the succeeding generation was generated based on the previous one. Each individual in the previous generation might have influence on the current

generation. The size of genetic samples was around 40 – compare this size with (King et al. 2000). Amount of analyzed individuals can make the difference in outcomes (Fig. 1). Cut-off values of statistics $\ln\hat{\beta}_1$ and $\ln\hat{\beta}_2$, were determined by the 0.05 percentile of the empirical distributions. The simulation included creation of 100 unlinked histories with constant amount of individuals N = 20 000. Each individual had 30 microsatellites, and 100 samples were taken from every generation with a number divisible by 100 000. For each history, 1 000 000 generations were simulated with mutation rate $v = 5 \times 10^{-4}$.

For simulation there was used co-designed by the author software called GenSim. The software was written in *C#* programming language in .NET framework, using the Mersenne Twister random number generator. The training sets obtained from simulations were used by one layer and two layers perceptrons which served as the models for the new ANN-based test (Cyran and Myszor 2008b, 2008c). The perceptrons were utilized because these networks are known to be universal tools for approximation problems, contrary for example to probabilistic neural networks which learn much faster but are dedicated primarily for classification. This issue is further discussed after presenting details of the simulation model and its results.

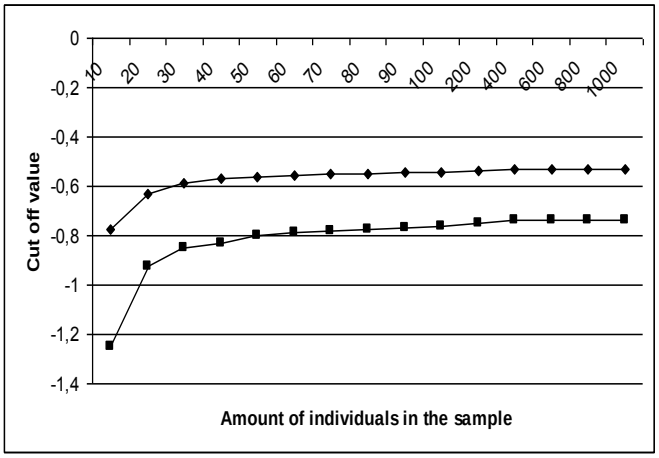


Fig. 5.2:1. Cut-off values of statistics $\ln\hat{\beta}_1$ ♦ and $\ln\hat{\beta}_2$ ■
Rys. 5.2:1. Wartości odcięcia statystyk $\ln\hat{\beta}_1$ ♦ and $\ln\hat{\beta}_2$ ■

To make the experiments as close to reality as possible, the samples of n individuals were taken from populations, and each sample contained fewer members than the whole population, what is the case in studying actual populations. The algorithm of time-forward simulation consists of the following steps:

- Preparation of initial population composed of N individuals. All individuals have the same amount of unlinked microsatellites. This step includes initialisation of each

microsatellite with the same value. Unless the goal is simulation of vanishing microsatellites, the initial size should be properly high.

- Run the simulation for 2*N* to 4*N* iterations in order to reach mutation drift equilibrium (Donnelly et al. 2001) and obtain a sample resembling an actual one.
- During each iteration creation of the next generation of *p* individuals (*p* is determined by assumed changes of population size).
- For each member of the new generation, the parent in the previous generation is drawn, microsatellites from the parent are chosen, and for each mutations are applied according to SSMM model (one parent can have many children).
- Creation of as many generations as needed.

As the statistical information about a population the growth coefficient based on microsatellites, called the imbalance index was computed. There are two estimators of imbalance index, Kimmel's estimator

$$\ln \hat{\beta}_1 = \ln \hat{\theta}_{\bar{V}} - \ln \hat{\theta}_{\bar{P}o} \tag{5.2:1}$$

and King's and Kimmel's estimator

$$\ln \hat{\beta}_2 = \frac{1}{m}\sum_{i=1}^{m}\left((\ln \hat{\theta}_V)_i - (\ln \hat{\theta}_{Po})_i\right). \tag{5.2:2}$$

In the above formulae, *m* is the amount of microsatellites, $\hat{\theta}_v$ denotes the allele size variance estimator of the composite parameter $\theta=4N\mu$, which is connected with the scale of the process, and $\hat{\theta}_{Po}$ denotes the homozygosity estimator of $\theta=4N\mu$.

Moreover, the variance estimator for a given microsatellite is given by

$$\hat{V} = \frac{1}{n(n-1)}\sum_{i\neq j}(X_i - X_j)^2 = \frac{2}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2, \tag{5.2:3}$$

where *n* is the amount of individuals in the sample, X_i is the length of a microsatellite of the i^{th} individual, and $\overline{X}$ is the mean of the length of microsatellites among individuals. The variance estimator across microsatellite loci is

$$\overline{V} = \frac{1}{m}\sum_{i=1}^{m}\hat{V}_i \cdot \tag{5.2:4}$$

Finally,

$$\hat{P}_0 = \frac{\left(n\sum_{k\in K} p_k^2 - 1\right)}{n-1}, \tag{5.2:5}$$

where *K* is a set of allele length in the sample, and

$$p_k = \frac{n_k}{n}, \tag{5.2:6}$$

with n_k denoting the amount of alleles with length equal to k. Averaging (5) across microsatellites yields

$$\overline{P_0} = \frac{1}{m}\sum_{i=1}^{m} \hat{P}_{0i}, \tag{5.2:7}$$

and based on (5) or (7) the homozygosity estimator of the composite parameter is computed from

$$\hat{\theta}_{Po} = \frac{1/\hat{P}_0^2 - 1}{2}, \tag{5.2:8}$$

which is plugged to (2) or (1), respectively.

The reader interested in more in depth understanding of imbalance indices should refer to (King et al. 2000) where these equations are explained in detail. In particular, the characteristics of these estimators are there described, based on a series of samples for populations undergoing growth of different types and magnitudes. Simulations described by King et al. (2000) were based on coalescent methods, and these simulations were repeated by Cyran and Myszor (2008a) using forward-time simulation method.

The correlation between both estimators is presented in Fig. 2, which presents the results for 100 unlinked histories, each containing $N = 2500$ individuals with 30 microsatellites mutating with a rate $v = 5 \times 10^{-4}$. After simulating 100 000 generations, 100 samples containing 40 individuals were taken from the population. For each population mean of $\ln \hat{\beta}_1$ and $\ln \hat{\beta}_2$ were computed from these 100 samples. Those means were put on graph in Fig. 2.

At the beginning of mentioned simulations all microsatellites had the same length, and then the simulation of $2N$ to $4N$ generations was started in an initialisation process (Donnelly 2001). During this initial time period the values of estimators are stabilising and populations reach mutation – drift equilibrium. After this pre simulation period, it is possible to take significant samples from a population and simulate a population growth.

An important issue in the forward-time computer simulation is the minimal amount of unlinked histories that is needed in order to gain significant results. Every unlinked history has different values of imbalance index estimators and empirical tests showed that for constant samples of different sizes, around 60 histories were enough to achieve stabilization of imbalance index estimators' cut-off values (Fig. 3).

In the experiments, two typical types of population growth were simulated in forward-time:

a) Exponential growth from N = 2 500 individuals to 5 000, 25 000 and 250 000. For the same final population size, the rate of exponential growth varied because of different

times of achieving final population size. We used up to 11 different time scales to simulate growths, from as fast as lasting only 625 generations, to as slow as lasting even 640 000 generations. Unique connection of final population size and time of reaching final population size is described in what follows as a scenario.

b) Stepwise growth from N = 2,500 individuals to 5,000, 25,000 and 250,000.

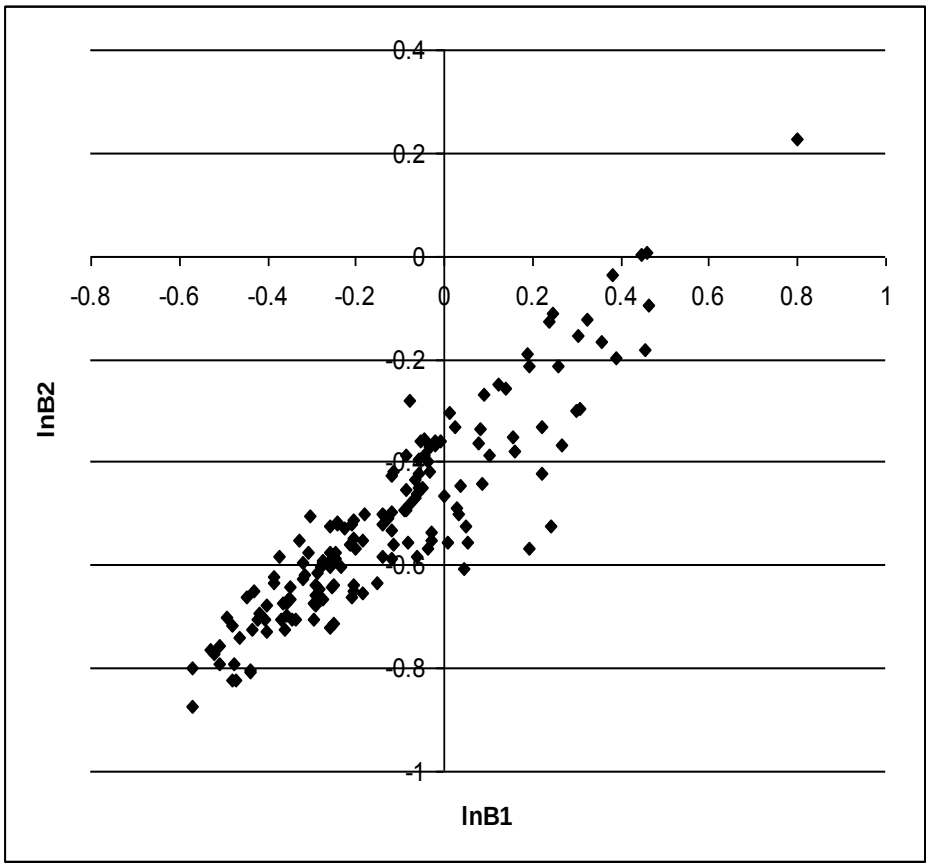


Fig. 5.2:2. Estimator $\ln \hat{\beta}_2$ as a function of $\ln \hat{\beta}_1$
Rys. 5.2:2. Estymator $\ln \hat{\beta}_2$ jako funckcja $\ln \hat{\beta}_1$

For each population growth scenario (if not said different in specific experiment description) there were created 100 independent histories, and 100 samples were drawn from the final generation, each sample containing 40 individuals (one individual couldn't be found twice in one sample, but might be found in several different samples). In each individual we simulated evolution of 30 unlinked microsatellites, and, as it was mentioned, the mutation rate was set at $\nu = 5 \times 10^{-4}$.

Interestingly, as a result of simulations, different cut-off values of imbalance index estimators were obtained, as compared to those reported by King et al. (2000). Powers of estimators for new cut-off values (for $\ln \hat{\beta}_1$ cut-off value was equal – 0.51, and for $\ln \hat{\beta}_2$ it was equal – 0.79) are lower than those estimated by King et al. (2000) for cut-off values – 0.32

and -0.65, respectively. These simulation results are visible in Fig. 4 and Fig. 5, for exponential and step-wise growths, respectively. Simulations were performed for the population, whose size started from $N = 2{,}500$ and changed to $N = 5{,}000$ (a), $N = 25{,}000$ (b) and $N = 250{,}000$ (c) individuals, respectively. The horizontal axis represents the number of generations after which the estimators are computed. Note, that in the case of exponential growth, these numbers correspond also to the duration of the growth.

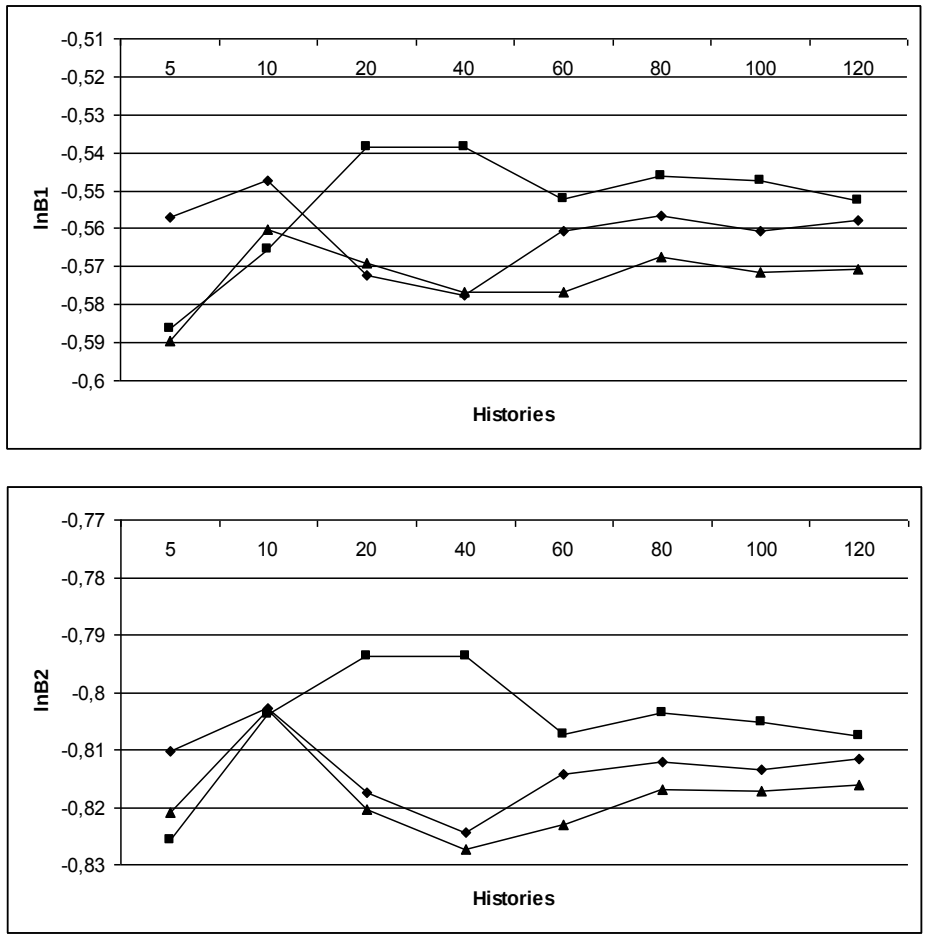
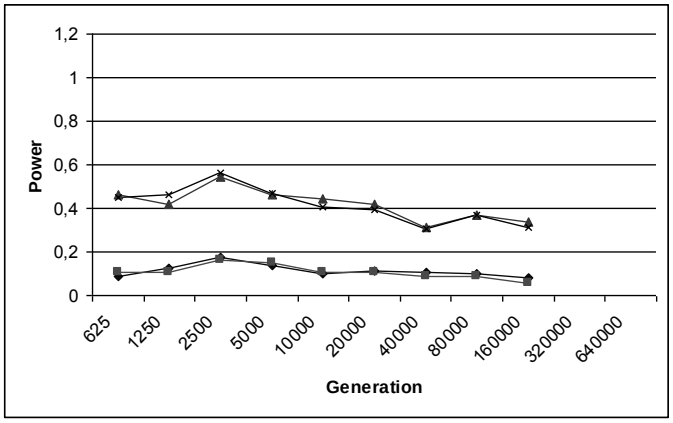


Fig. 5.2:3. Cut-off values of $\ln \hat{\beta}_1$ (a) and $\ln \hat{\beta}_2$ (b) based upon population with constant size of 2,500 (♦) 5,000 (■) and 20,000 (▲) individuals

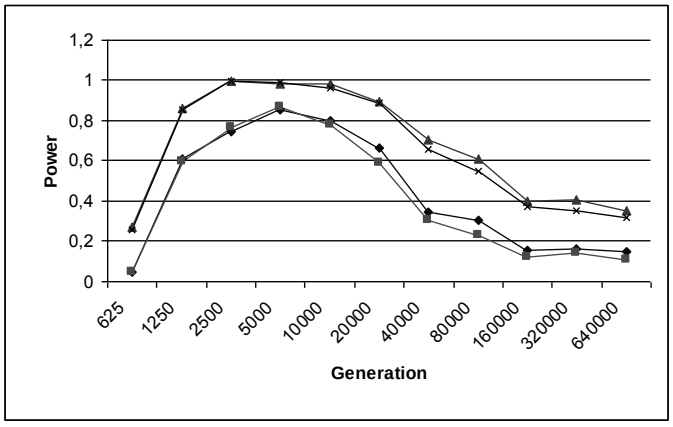
Rys. 5.2:3. Wartości odcięcia $\ln \hat{\beta}_1$ (a) oraz $\ln \hat{\beta}_2$ (b) dla populacji ze stałym rozmiarem liczące 2,500 (♦) 5,000 (■) oraz 20,000 (▲) osobników

It should be stressed that all conditions of experiments were the same as in (King et al. 2000) and the only difference was different method of simulation used in the experiments, namely the forward-time simulation, which gives more reliable results. Since forward-time simulation methods are closer to real life scenario it might be appropriate to consider using this method of simulation, especially having in mind the increase in computational power of computers and possible parallelism which can be implemented in simulation algorithms.

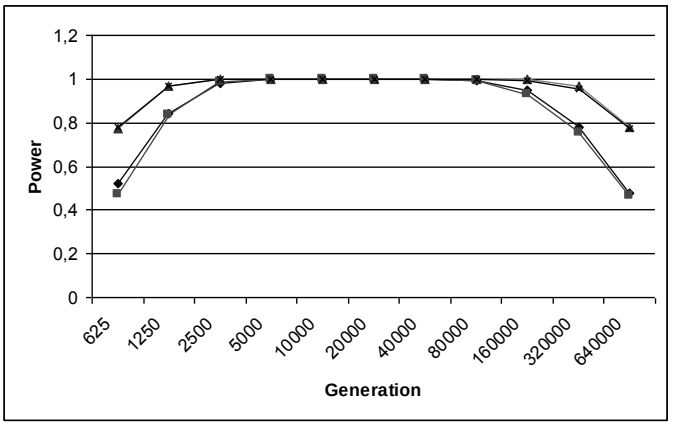
Moreover, the forward-time simulations are applicable for arbitrary complex demographic histories, including geographic structure of the population and time-inhomogeneous reproduction schemes.



(a)



(b)



(c)

Fig. 5.2:4. Power of $\ln \hat{\beta}_1$ (▲) and $\ln \hat{\beta}_2$ (*) based on coalescent methods, and $\ln \hat{\beta}_1$ (♦) and $\ln \hat{\beta}_2$ (■) based on time-forward computer simulation, for exponential growths

Rys. 5.2:4. Moc $\ln \hat{\beta}_1$ (▲) i $\ln \hat{\beta}_2$ (*) na podstawie koalecentu, oraz $\ln \hat{\beta}_1$ (♦) i $\ln \hat{\beta}_2$ (■) na podstawie symulacji komputerowych w przód dla wzrostu wykładniczego

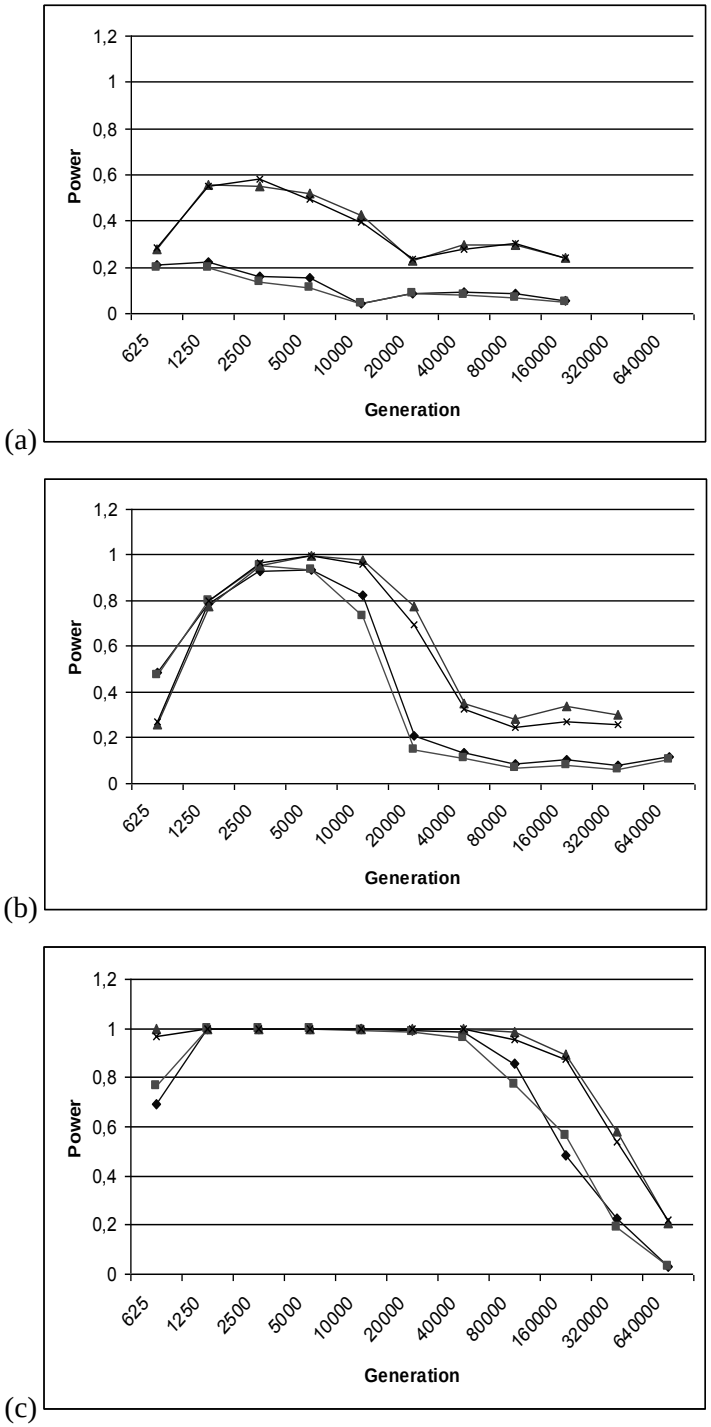


Fig. 5.2:5. Power of $\ln \hat{\beta}_1$ (▲) and $\ln \hat{\beta}_2$ (*) based on coalescent methods, and $\ln \hat{\beta}_1$ (♦) and $\ln \hat{\beta}_2$ (■) based on time-forward computer simulation, for step-wise growths

Rys. 5.2:5. Moc $\ln \hat{\beta}_1$ (▲) i $\ln \hat{\beta}_2$ (*) na podstawie koalecentu, oraz $\ln \hat{\beta}_1$ (♦) i $\ln \hat{\beta}_2$ (■) na podstawie symulacji komputerowych w przód dla wzrostu skokowego

Forward-time computer simulations were repeated sufficiently many times in order to create samples used as training data for artificial neural networks. As stated before, the

Wright-Fisher model was used with provided dynamic description of the demographic evolution to allow changes between amounts of individuals in the population. Each simulation comprised at least *8N* = 20 000 initializing generations (the conservative choice of that number to be 8*N* assured achieving the mutation-drift equilibrium – see section 3.3) before the valid simulations of growth started.

For the ten-fold stepwise growth from 2,500 to 25,000 individuals, there were created a number of samples with many combinations of simulation parameters values, such as: mutation rate (2.5×10^{-4}, 5×10^{-4}, 7.5×10^{-4}) amount of individual's microsatellites (10, 30, 40), amount of individuals in examined sample (10, 40, 70). Based on these simulations, the power of ANN-based test was computed and compared with the power of $\ln\hat{\beta}_1$ for the same samples.

In the study, there were used one and two-layers feed-forward neural networks with sigmoid neurons, whose network excitation *n* is given by (2.2:1), and the output signal *y* by (2.2:2). Appropriate size of network (amount of layers and neurons) is an important feature, which should be considered during design of the ANN. If the network is too small it might be unable to achieve desired global error value during learning. On the other hand, networks with too many neurons or layers might remember how to recognize all learning samples and lost generalization capabilities (see section 2.2.1 for details). It is also important to choose correct learning examples. Ideally, the samples in the learning set should cover uniformly all relevant for the problem cases.

The inputs for the first layer are fed by some relevant for the problem, population genetics-based statistics, after normalization. These statistics include Kimmel's estimator of imbalance index given by (1), King's and Kimmel's estimator of imbalance index given by (2) as well as two tests defined by Reich et al. (1999), the inter locus estimator *g* and the within locus estimator *k*.

The inter locus estimator *g* is the ratio of the observed and predicted variances of the allele length

$$g = \frac{Var(\hat{V})}{\frac{4}{3}\bar{\hat{V}}^2 + \frac{1}{6}\bar{\hat{V}}}. \tag{5.2:9}$$

Observe that in the above formula the numerator denotes the observed variance of the allele length given by

$$ObservedVariance(\hat{\theta}_V) = Var(\hat{V}) = \frac{1}{n-1}\sum_{j=1}^{m}\left(\hat{V}_j - \bar{\hat{V}}\right)^2, \tag{5.2:10}$$

and the denominator has got the meaning of the variance value predicted in the drift-mutation equilibrium

$$ExpectedVariance(\hat{\theta}_V) = \frac{4}{3}\bar{\hat{V}}^2 + \frac{1}{6}\bar{\hat{V}}. \quad (5.2:11)$$

In the above formulas $\hat{V}_j$ is an unbiased estimator of variance of the allele length distribution at locus *j*, and $\bar{\hat{V}}$ is the mean of the unbiased estimators of variance of allele length distributions. More detailed description of these equations the reader can find in (Reich et al. 1999).

The within locus estimator *k* is given by

$$k = 2{,}5 * Sig^4 + 0{,}28 * S^2 - 0{,}95/n - Gam_4, \quad (5.2:12)$$

where

$$Sig^4 = \frac{(n^2 - 3n + 3)}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2 - \frac{1}{(n-2)(n-3)}\sum(X_i - \overline{X})^4, \quad (5.2:13)$$

$$Gam_4 = \frac{(n^2 - 2n + 3)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}(X_i - \overline{X})^4 - \frac{(6n-9)}{n(n-1)(n-2)(n-3)}\left(\sum_{i=1}^{n}(Xi - \overline{X})^2)\right)^2, \quad (5.2:14)$$

and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2. \quad (5.2:15)$$

In the above equations, described in detail in (Reich and Goldstein 1998), S^2 has got the meaning of an unbiased estimator of the variance, Sig^4 is an unbiased estimator for the variance squared, and Gam_4 is the fourth central moment of the allele length distribution.

Estimators of the imbalance indices and values of the inter locus estimator *g* usually receive values from a range 0 to 1 so it is possible to put them directly on the network input. The value of the within locus coefficient *k* should be divided by a number of microsatellites *n* to obtain normalized version, which can be the input for the ANN. These statistics are designed to detect different histories of samples drawn from populations with constant size (Fig. 6a) and those which underwent in past a substantial growth (Fig. 6b) – the use of them in a one common properly designed test should give a power, which is greater than the power of any of these tests when used separately.

The qualitatively visible difference in lengths of the branches leading from the most recent common ancestor in both genealogies presented in Fig. 6 is the reason why the distributions of the length of alleles are also different. For the constant population size the old branches are long and therefore they accumulate a lot of mutations what is reflected in two or

three-modal distributions of the allele length (Fig. 7). This is not so for the population evolving after significant growth. The corresponding genealogy has got short branches leading from the most common ancestor, so the mutations accumulate in young branches yielding unimodal distributions (Fig. 8).

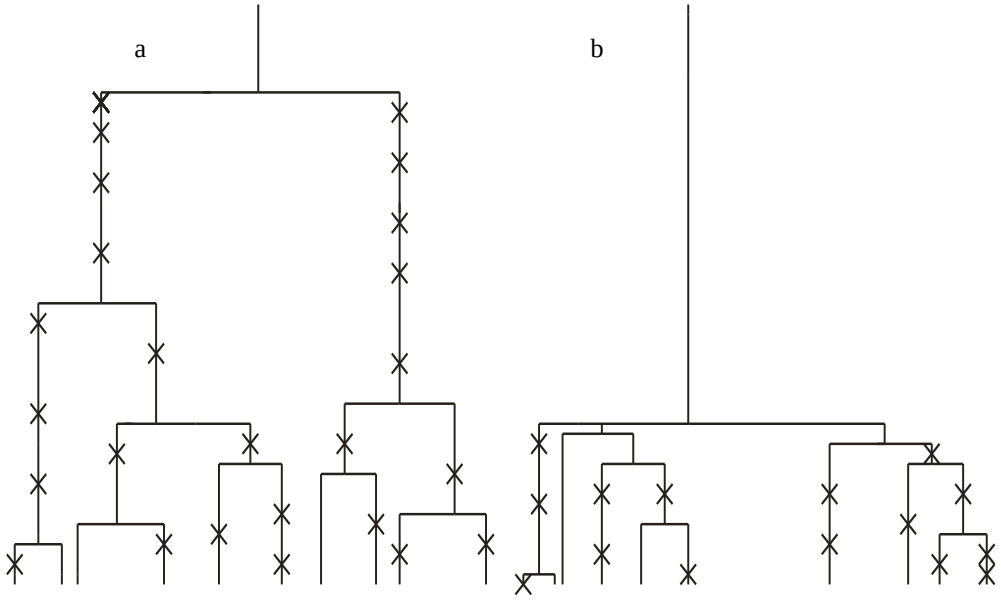


Fig. 5.2:6. Genealogies with mutations (crosses) of 10 individuals from a population with present size 20,000. (a) constant population size (b) 100-fold growth 8,000 generations ago
Rys. 5.2:6. Genealogie z mutacjami (krzyżyki) 10 osobników z populacji o końcowym rozmiarze 20,000. (a) stały rozmiar populacji (b) 100-krotny wzrost 8,000 pokoleń temu

The output of the network was the normalized value of the test with the experimentally determined critical value at given significance level. Intuitively, the greater value of the network output (minimum value is 0, maximum value is 1), the bigger probability that the population experienced expansion. The learning set contained similar numbers of samples from constant and from growing populations. Samples from growing populations came from populations that underwent the stepwise growth and the exponential growth.

When the network learning was finished the critical cut-off value was determined as 0.95 percentile of the output values generated for histories with a constant population size. The set of examples which we used to estimate a power of the ANN-based test comprised the training set, but this training set was just a little fraction (about 5%) of all samples used in estimation of the power.

Interestingly, the greatest power was obtained using single layer ANN containing only one output neuron. For such a simple structure there is a possibility to give the explicit equation of the trained network, and therefore to define analytically a new test γ (with experimentally obtained weights) given as

$$\gamma = \left\{1 + \exp\left(4.201 \ln \hat{\beta}_1 + 2.417 \ln \hat{\beta}_2 + 3.842 \frac{k}{n} - 1.247\, g\ + 1.511 \right)\right\}^{-1}. \qquad (5.2{:}16)$$

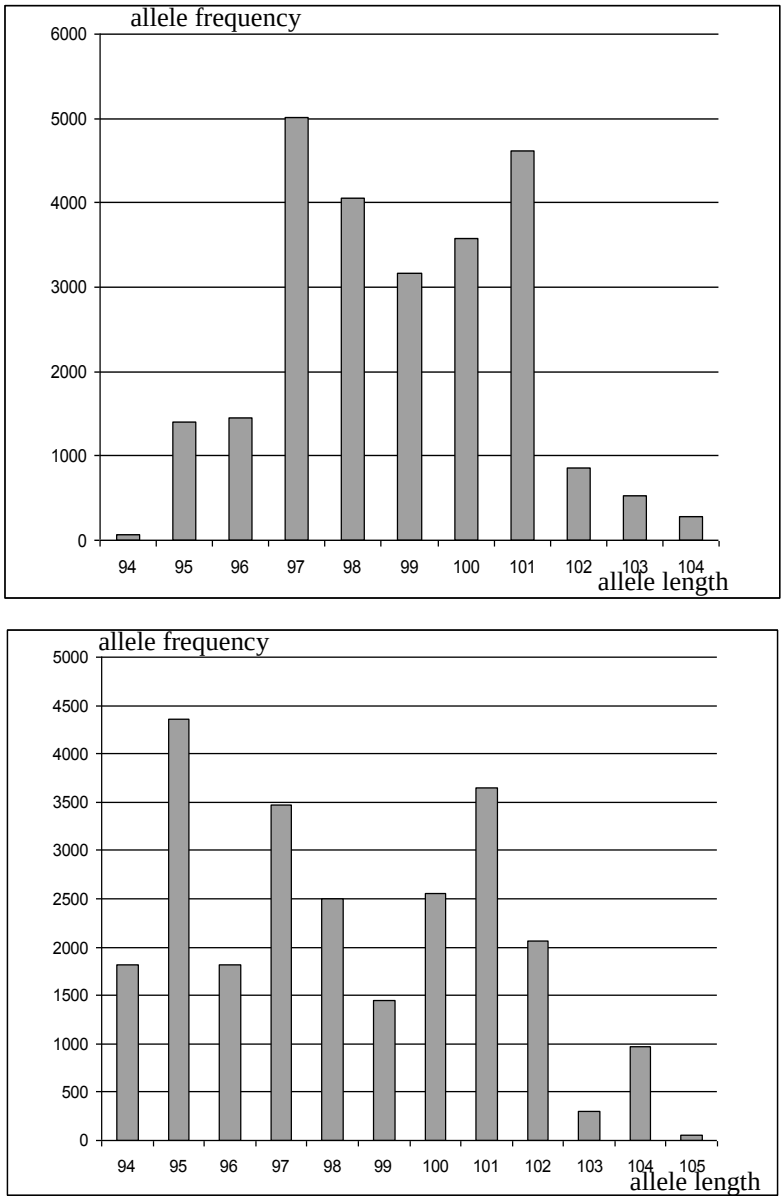


Fig. 5.2:7. Histograms of the allele length for constant population size of 25,000 individuals with typical for microsatellite mutation rate equal 5×10^{-4}

Rys. 5.2:7. Histogramy długości alleli dla stałego rozmiaru populacji 25,000 osobników z typowym dla mikrosatelitów tempem mutacji równym 5×10^{-4}

Test γ returns values from a range (0,1) with the critical cut-off value equal to 0.797 at a significance level 0.05 (if the test returns greater value we assume that sample comes from a population that experienced growth). The power of γ was compared with powers of tests belonging to the most powerful growth detectors, namely the estimators of imbalance index.

Based on empirical distribution of imbalance index estimators values for constant population size, the critical values for these tests are estimated to be equal – 0.51, and – 0.787 for $\ln \hat{\beta}_1$ and $\ln \hat{\beta}_2$, respectively.

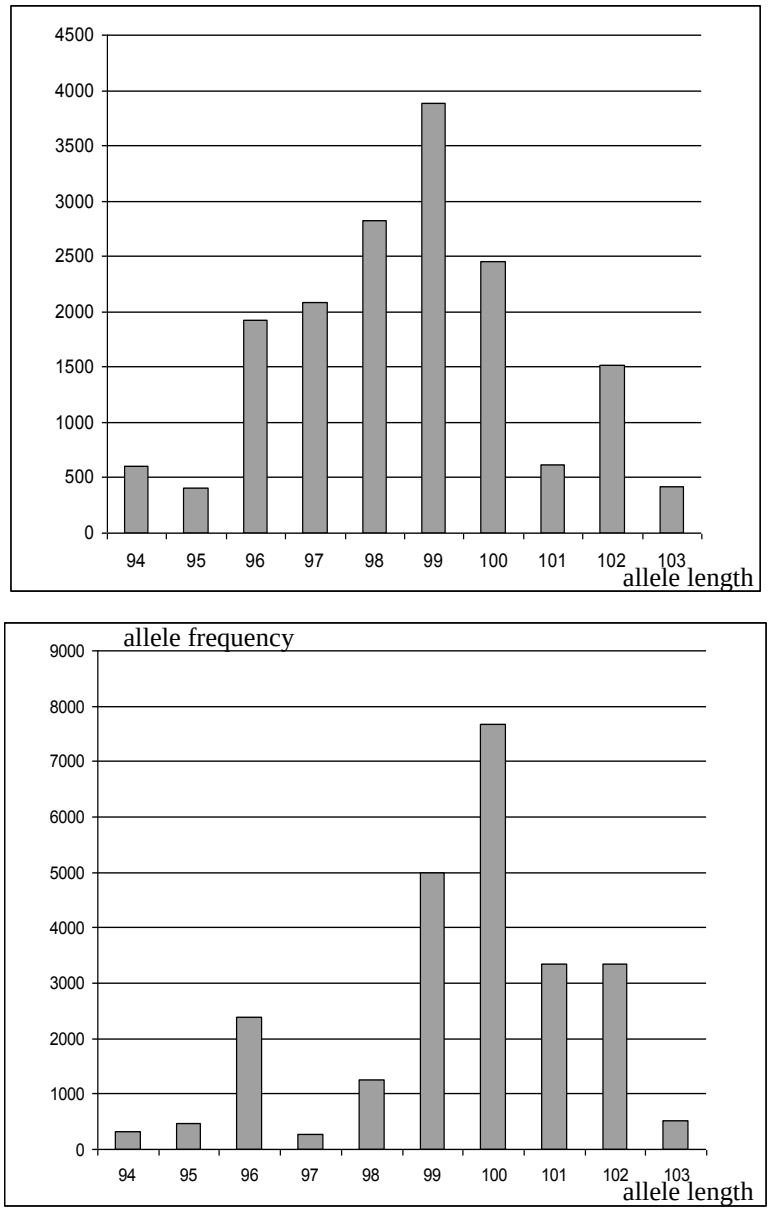


Fig. 5.2:8. Histograms of the allele length for population of the size 25,000 individuals which underwent 10-fold increase 20,000 generations ago ($\mu = 5 \times 10^{-4}$)
Rys. 5.2:8. Histogramy długości alleli dla populacji o rozmiarze 25,000 osobników, która doświadczyła 10-krotnego wzrostu 20,000 pokoleń temu ($\mu = 5 \times 10^{-4}$)

To obtain these values, there were created 150 histories of constant population size with $N = 2\,500$ individuals, which were simulated for 100 000 generations. Starting from the 50 000th generation, the samples from every generation divisible by 10 000 were taken. From each history 100 samples were analysed, each containing 40 individuals. For small stepwise

growth (like two-fold growth) there is no difference in a power between estimators of imbalance index and the test γ (Fig 9a) – actually all tests have low power for such small growth. However, for ten-fold growth, there is a visible difference in power of tests (Fig 9b). The ANN-based test γ is able to detect growth earlier than methods based on imbalance index, and moreover, the signal about an expansion stays longer. For 100-fold growth (Fig 9c), test γ detects the growth even earlier and for longer time.

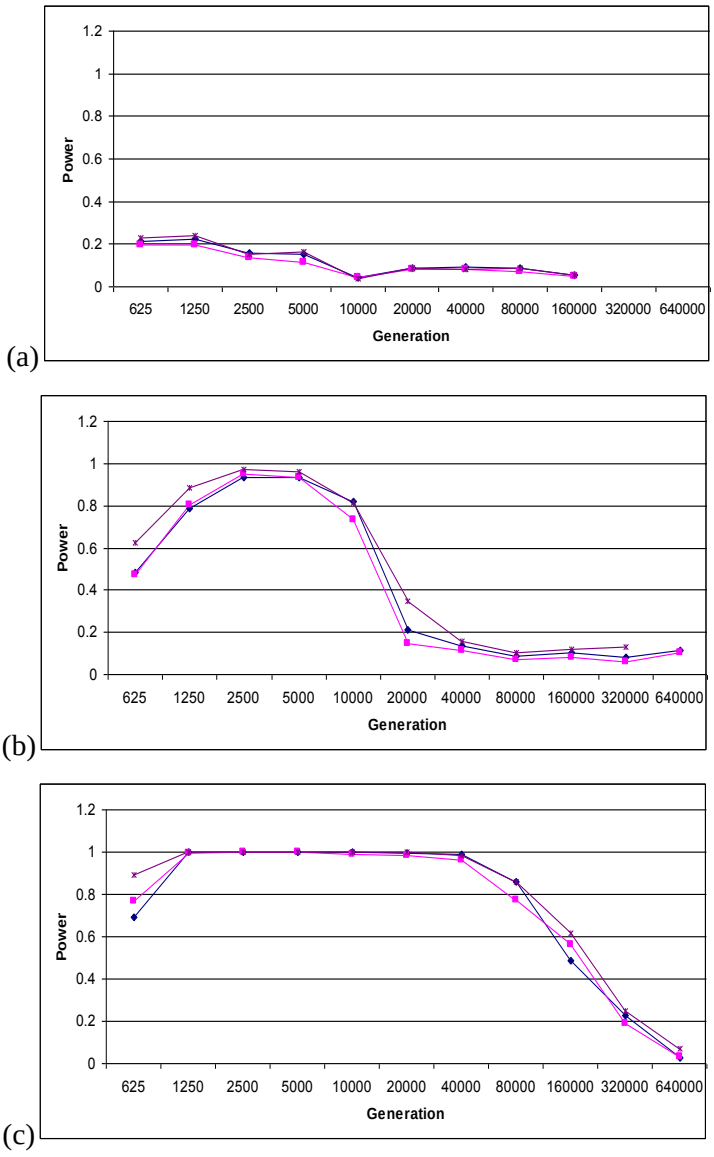


Fig. 5.2:9. Powers of $\ln\hat{\beta}_1$ (♦), $\ln\hat{\beta}_2$ (■) and γ (*) tests. Populations experienced stepwise growth from N = 2 500 to (a) 5,000, (b) 25,000 and (c) 250,000 individuals

Rys. 5.2:9. Moce testów $\ln\hat{\beta}_1$ (♦), $\ln\hat{\beta}_2$ (■) i γ (*). Populacje doświadczyły skokowego wzrostu z N = 2 500 do (a) 5,000, (b) 25,000 i (c) 250,000 osobników

Figure 10 presents the results for the exponential growth spread over different time periods. For each generation number marked on the graph, there was created a set of 100 unlinked histories in which the final population size was reached at the marked time. For

small exponential growths, powers of all tests are low (Fig. 10a). In the case of greater exponential expansion, test γ has greater power and might detect growth for longer time (Fig. 10 b and c). Fig. 10 shows that test γ usually give outcomes better than other available tests based on microsatellites.

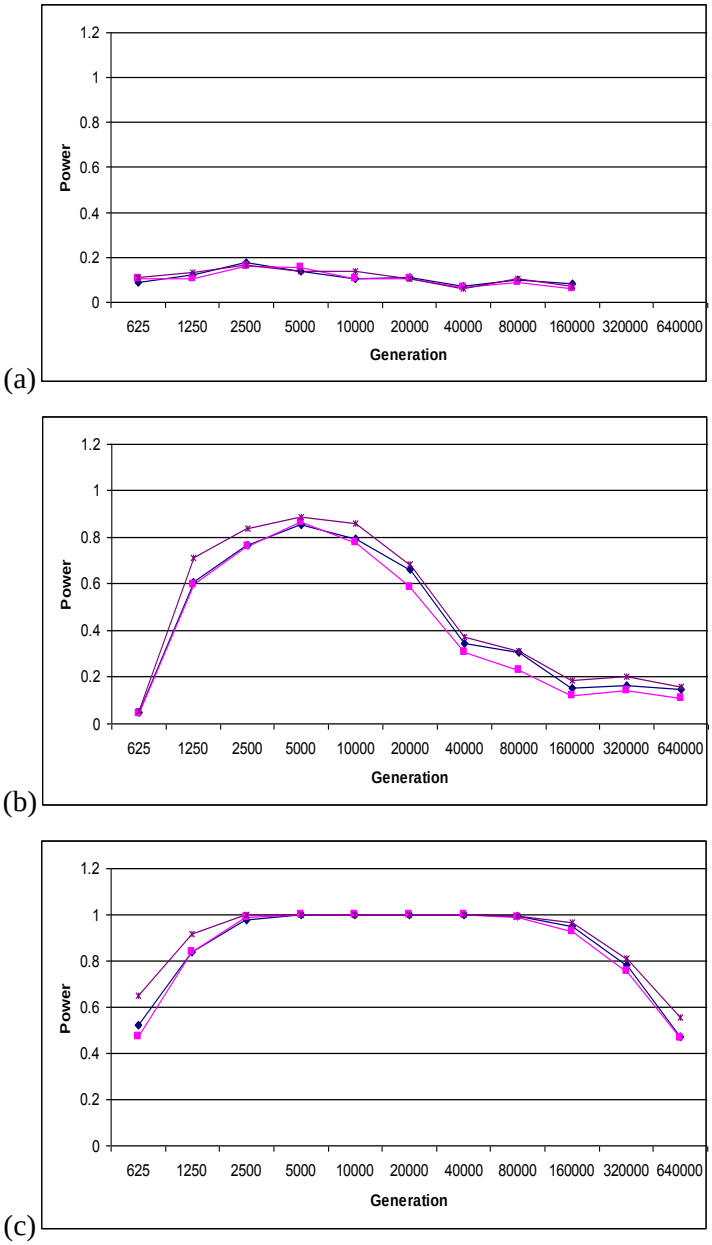


Fig. 5.2:10. Powers of $\ln \hat{\beta}_1$ (♦), $\ln \hat{\beta}_2$ (■) and γ (*) tests. Populations experienced exponential growth from N = 2 500 to (a) 5,000, (b) 25,000 and (c) 250,000 individuals

Rys. 5.2:10. Moce testów $\ln \hat{\beta}_1$ (♦), $\ln \hat{\beta}_2$ (■) i γ (*). Populacje doświadczyły wykładniczego wzrostu z N = 2 500 do (a) 5,000, (b) 25,000 i (c) 250,000 osobników

Differences in a power of tests are especially visible for populations undergoing growths with bigger rate (i.e., such which are potentially detectable). In Fig. 11 it is visible that for small growths (left side of the graph, little difference in amounts of individuals between two

generations) powers of $\ln \hat{\beta}_1$ and γ are similar but for bigger growths (right side of the graph) test γ has a greater power. The values of power are counted for the number of generations marked on horizontal axis.

In the above study it was demonstrated that the properly trained neural network defines a novel statistical test γ given by (16) which has greater power in the detection of population growth, than any other tests based on microsatellites, as it is showed in Fig. 9, 10, and 11. It is easy to understand taking in mind that the test γ uses the information involved in other tests and the importance of information in any particular test is weighted by the neural network according to the rule learned from training data obtained from extensive forward-time computer simulations.

It was proved by King et al. (2000) that the power of the imbalance indices $\ln \hat{\beta}_1$ and $\ln \hat{\beta}_2$ is greater than that of *k* and *g* statistics defined by Reich and Goldstein (1998) and Reich et al. (1999), however the design of the γ test showed that additional information covered in these two latter tests can further increase the power of resulting statistic.

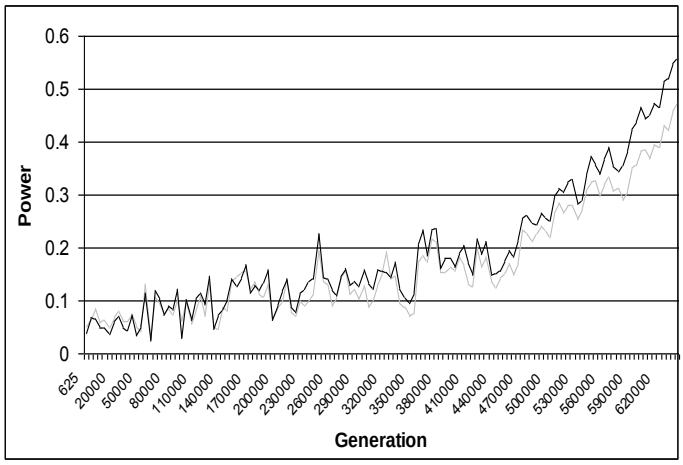


Fig. 5.2:11. Power of γ (black) and $\ln \hat{\beta}_2$ (gray) for population which undergoes exponential growth from *N* = 2,500 to 250,000 individuals during 640,000 generations

Rys. 5.2:11. Moc γ (czarny) oraz $\ln \hat{\beta}_2$ (szary) dla populacji która doświadczyła wzrostu wykładniczego z *N* = 2,500 do 250,000 osobników w czasie 640,000 pokoleń

5.3. Mitochondrial Eve dating – robustness of the Wright-Fisher model

In this section, there are considered three different models for calculating the distribution of the time to coalescence of a pair of alleles used for dating Mitochondrial Eve period. Comparison of these models allows to answer the question of how relevant for the model expectations are departures from panmictic population (in the case of the Wright-Fisher

model) and from the assumption about large size of population (in the case of the coalescent method used for populations with visible stochastic effects).

These three models include the Wright-Fisher model with discrete generations (see section 3.2), the coalescent-based method with continuous time scaled by variable in time size of population (described in section 3.5), and the O'Connell limiting model dedicated to branching processes (described in section 3.6). The choice of the Wright-Fisher model and the coalescent methods is evident having in mind their popularity. Why, the less common O'Connell model was also used for comparison, requires some justification. While the details for these reasons are given in sections 3.6 and 3.7, they can be summarized here by mentioning the independence of the model of the shape of the offspring distribution with the same expected value and bounded variance.

All three mentioned models are applied for stochastic population growth approximated by a slightly supercritical Galton-Watson branching process. To be able to compare theses methodologies reliably, there is designed a computational framework for estimation of the two-allele coalescence distribution in any of these models as well as in a model based on full record of the population history, and therefore giving opportunity to compute precisely desired parameters conditional on simulated genealogy. Having simulated several thousand genealogies it is possible to estimate parameters unconditionally with a great accuracy.

There could be some doubt whether the use of the time to coalescence of two alleles is an adequate tool in genealogical applications. To answer the problem there could be also considered the problem of coalescence of a sample of n alleles randomly chosen from a population. However, the nature of the recursion intrinsically involved in it as well as difficulty with association of the results with known genetic indices make the use of it troublesome. Therefore, although perhaps there is a considerable room for possible applications, the analysis of aforementioned problem will not be discussed here because of difficulties with association of such distributions with genetic data. Therefore, the main reason of why a distribution of a time to coalescence of a pair of alleles is used in this study, is the ease of association of its expected value with the average pairwise mutation difference between two randomly chosen individuals. These two notions must be only scaled by the mutation rate to make it possible to estimate one from the other.

The ease of aforementioned associations is in a clear contrast to samples composed of more than two alleles, analysis of which requires complex phylogenetic methods. In that latter case the problem is caused by various, and in majority of cases unknown trees, relating the individuals in a sample. Phylogenetic methods attempt to use all genetic information covered in a sample to build the genealogy of the sample (e.g. Griffiths and Tavaré 1995) and there exist some computer programs such as Griffiths' *genetree* for inferring the phylogenetic trees. While these methods are often used and tend to give estimates with smaller variance

than those based on a pairwise differences, serious difficulty with comparison of them with the O'Connell model serving as a standard excluded them from more detailed consideration.

To answer the question about the sensitivity of the distribution of the time to coalescence to departures from the Wright-Fisher and the coalescent models two approaches are basically possible. The first approach requires storing the whole simulated genealogy of a population undergoing periods with visible stochastic effects and evolving according to arbitrary non-multinomial sampling scheme. Then, by averaging over genealogies, the experimental distribution of the times to the coalescence can be found and compared to that obtained in the Wright-Fisher and the coalescent models. It is a very general approach not limited to any generation-to-generation sampling scheme and assumption about large population size.

In particular it can be applied for arbitrary progeny distribution (possibly changing in time) used to model the evolution of the population as a branching process, whose beginning represents period with clearly visible stochasticity. However, one can argue that this approach has strong limitations in the number of generations it can model. Except for small population sizes, it requires large amount of memory for storing information about each generation, and therefore it looks practically not feasible for simulations of number of generations required for dating the Mitochondrial Eve. Since the interest of this study is in examining robustness to departures from the model assumptions for long-term histories of human population, which assumes multiple repeats of simulations, this approach at first seems not to be feasible.

In the alternative approach, the population history is simulated and only the time course of its size is recorded. Assuming the offspring distributions other than Poisson and simulating population undergoing period of small size (thus generating departures from the Wright-Fisher model and the coalescent models approximated by diffusion model) we can compute the coalescence distribution in the Wright-Fisher and the coalescent models. Such coalescence distributions can be then compared to a coalescence distribution obtained for the same population history in some other model which would be treated as a standard. Considering the O'Connell (1995) model as a standard is dictated by the fact that in this model it is possible to calculate the interesting distributions of the time to coalescence independently of the sampling scheme and variance of offspring distribution, if only the population evolves as a Markov slightly supercritical branching process.

More precisely, the O'Connell model is independent asymptotically of the shape of the progeny distribution for given mean, as long as the variance of the distribution is bounded. The results of experiments verifying this fact were reported by Cyran and Kimmel (2004a, 2004b) and Cyran (2007b, 2007d). They were also used in Cyran and Kimmel (2005) for conservative estimation of parameter α (see section 3.6 for the definition of this parameter influencing the expansion rate of the population) in a problem of hypothetical Neanderthal admixture to modern human mitochondrial DNA gene pool (this problem is discussed in

detail in the section 5.4). However, this methodology lacks one important feature which could be taken into consideration only in the first approach. Namely, having only the sizes of the population and lacking its full genealogy it is impossible to distinguish between the time of whole simulation started from one individual, and that elapsed from the MRCA. The problem becomes visible if one imagines that founder of the process, definitely being the common ancestor of the population evolved, not necessarily (and in fact rarely) is the *most recent* common ancestor. Having no possibility to distinguish between the two, it was assumed in earlier studies (Cyran and Kimmel 2004a, 2004b, 2005) that the time between the founder and the MRCA is relatively short as compared to the time of the whole process. Therefore, both times were treated as identical, having no information to what extent this simplification can be justified.

With the increase of computational power of computers and the capacities of the memories it was possible to return to the problem by implementing the first approach indirectly, i.e. with the help of the O'Connell model, after experimental verification that it is feasible to simulate and record full genealogies for such number of generations for which the validity of the O'Connell model asymptotic results is clearly true. The author developed software capable for simulating full genealogies of at least 10^2 generations under arbitrarily chosen distribution of offspring and with parameters of the branching process identical to those which could reflect the long-time (i.e. for about 10^4 generations) evolution of modern humans.

When the variance of the offspring distribution is small (resulting in smaller population sizes, given the identical mean) it was also possible to simulate and store full genealogies for 10^3 generations, but for the larger variances it was still impossible, so it was verified whether the asymptotic properties of the O'Connell model hold for such small number of generations as 10^2. If it did not prove true, then basing on simulations of 10^2 generations it would not be possible to draw conclusions about the relative distance in time between the founder of the process and the MRCA of the population in the evolution comprising 10^3-10^4 generations.

However, if the asymptotic behavior of the slightly supercritical branching process was valid already for simulations comprising as little as 10^2 generations, the description of the evolution assuming the same parameters of the O'Connell model should become identical for simulations with any number of generations exceeding 10^2. This result is due to the fact that such demographies from the definition would resemble the limit model in greater detail than demographies having only 10^2 generations, but even for the latter the limit model holds. This allowed to use the O'Connell model as a theoretical standard extrapolating the full genealogy simulation results for arbitrary many generations after experimental verification that convergence to the asymptotic properties of the O'Connell distribution is sufficiently fast, and

therefore departures from the asymptotic behavior for more than 10^2 generations are negligible.

The comparison of the coalescence distributions in different models allows to observe how sensitive to departures from their assumptions is the estimate T_{MRCA} denoting the mtEve epoch. For this purpose, there was modeled the long-term demographic history of a population by the evolution of a Markov slightly supercritical branching process. A sample of n DNA sequences was considered, which was taken from such population with the average duration of a generation (in years) equal to λ.

Moreover, let us denote the average pairwise mutation difference in such sample by d_{avg} and the mutation rate per nucleotide per generation by μ. In the infinite sites model the genetic divergence rate between two species δ is equal to μ/λ so it is possible to estimate mutation rate using $\hat{\mu} = \delta\lambda$. Then, denoting the average time to the coalescence of two individuals in a population by T_{2c}, the expectation of d_{avg} is given by

$$E\left(d_{avg} \middle| K_0 = 1\right) = T_{MRCA}\hat{\mu}E\left(\frac{T_{2c}}{T_{MRCA}} \middle| K_0 = 1\right) \tag{5.3:1}$$

where K_0 is the number of those individuals at generation 0 whose descendants persist alive until present. Assuming that $T_{MRCA\_y} = \lambda\, T_{MRCA}$ is the equivalent of T_{MRCA} expressed in years, the moment based estimate for T_{MRCA_y} is

$$\hat{T}_{MRCA\_y} = \frac{\hat{d}_{avg}}{\delta E\left(\frac{T_{2c}}{T_{MRCA}} \middle| K_0 = 1\right)} \tag{5.3:2}$$

Apart from the O'Connell model, the expectation $E(T_{2c}/T \mid K_0{=}1)$ is obtained by performing computer simulation of the branching process starting from one individual, computing (according to the model specific methods) the empirical coalescence distribution conditional on the process, and then by calculating the required ratio of T_{2c} and T_{MRCA}. After simulation of several thousand processes the expectation of the ratio can be obtained. However, only in the model with the record of the full genealogy both times T_{2c} and T_{MRCA} are explicitly given within recorded genealogy and (2) can be applied directly. In others models only time T_{2c} can be computed and the time T_{MRCA} is not available directly. Instead, the time T i.e., the time of the process, is at the disposal. Certainly, the time T, being the time to the only individual initiating the branching process, is the time to the common ancestor of whole evolved population. However, as it was mentioned, rarely it is also the time to the *most recent* common ancestor because of the fact that many lineages of its direct and indirect progeny become extinct.

Nevertheless, it is possible to estimate the ratio of T_{MRCA} and T in simulations with fully recorded genealogy, and moreover, it is fortunate that limiting properties of the coalescence

distribution in the O'Connell model are valid for as little as 10^2 generations for which it is possible to perform simulations in the full genealogy model. In this way it is possible to point out in the O'Connell model T_{MRCA} relative to T and T_{2c} and using the limit theorem, it is possible to propagate this result to arbitrary many generations, in particular to the number of generations leading roughly to actual short-term effective female population size. This leads to the equation

$$\hat{T}_{MRCA\_y} = \frac{\hat{d}_{avg}}{\hat{\delta} E\left(\frac{T_{2c}}{T} \frac{T}{T_{MRCA}} \middle| K_0 = 1 \right)} = \frac{E\left(\frac{T_{MRCA}}{T} \middle| K_0 = 1 \right) \hat{d}_{avg}}{E\left(\frac{T_{2c}}{T} \middle| K_0 = 1 \right) \hat{\delta}}. \tag{5.3:3}$$

The estimates of variables δ and d_{avg}, necessary for estimate $\hat{T}_{MRCA\_y}$ according to (3), can be retrieved from genetic diversity data. Since the inbreeding effective population size is proportional to the variance of the offspring distribution, to demonstrate departures in both direction from the model's standard, Poisson (P) progeny distribution, apart from this distribution there was considered the binary fission (BF) distribution and the linear fractional (LF) distribution. The corresponding probability generating functions (pgfs) of these distributions are

$$f(s) = \sum_{k=0}^{\infty} s^k e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda + s\lambda}, \tag{5.3:4}$$

for the Poisson,

$$f(s) = p^2 + 2p(1-p)s + (1-p^2)s^2 = [p + (1-p)s]^2, \tag{5.3:5}$$

for the binary fission, and

$$f(s) = \frac{1-b-p}{1-p} + \sum_{k=1}^{\infty} s^k b p^{k-1} = 1 - \frac{b}{1-p} + \frac{bs}{1-ps}, \tag{5.3:6}$$

for the linear fractional distribution.

In the O'Connell model, since

$$E\left(\frac{T_{2c}}{T} \middle| K_0 = 1 \right) = \frac{1}{T} E\left[(T - D_T) \middle| K_0 = 1 \right] \tag{5.3:7}$$

and based on formula (3.6:28), the equation (3) becomes

$$\hat{T}_{MRCA\_y} = E\left(\frac{T_{MRCA}}{T} \middle| K_0 = 1 \right) \frac{\hat{d}_{avg}}{\hat{\delta} \left(1 - 2 \int_0^1 \frac{\hat{q}_r}{(1-\hat{q}_r)^2} (\hat{q}_r - 1 - \ln \hat{q}_r) dr \right)}. \tag{5.3:8}$$

where

$$\hat{q}_r = \frac{e^{-r\hat{\alpha}} - e^{-\hat{\alpha}}}{1 - e^{-\hat{\alpha}}}, \tag{5.3:9}$$

the expectation of the ratio T_{MRCA} and T should be taken from simulations with recorded full genealogies, and $\hat{x}$ denotes the estimate of the parameter x.

Therefore, to calculate from genetic variation data the MRCA epoch given by $\hat{T}_{MRCA\_y}$ the parameter $\hat{\alpha}$ is required. However, from simulation results concordant with the limiting properties of the O'Connell model it is possible to obtain the ratio $E(T_{MRCA} / T \mid K_0 = 1)$. Therefore, we can simultaneously estimate T_{MRCA_y} and α. From Theorem 3.6:4, equation (3.6:22), if Z_T is substituted as an estimate of its expected value, it follows that

$$Z_T = E\left(\frac{T}{T_{MRCA}} \middle| K_0 = 1\right) \frac{\sigma^2 \hat{T}_{MRCA\_y}}{2\lambda\hat{\alpha}} \left[\exp(\hat{\alpha}) - 1\right] \tag{5.3:10}$$

and estimates of T_{MRCA_y} and α are solutions of the system of equations (8) and (10), for given short-term inbreeding effective population size of females Z_T, and genetic data summarized by d_{avg} and δ.

The software designed by the author for simulation of branching processes in the context of its genealogy works in one of two modes. The first mode implements the full genealogy recording, thus allowing for explicit access for any desired feature of the model. In particular, it is possible to trace back the genealogy of a pair of individuals and to find their MRCA and therefore the actual time of coalescence. By random choice of a sample of, say 100 individuals, and determining the coalescence of the each pair in the sample (tracing all pairs in the whole population proved to be extremely time inefficient) it is possible to obtain, conditionally on the simulated tree, a histogram $H_{T2c|tree}$ of the times to the coalescence, which is the experimental approximation of the conditional coalescence distribution $P(T_{2c} = t \mid tree)$ in the full genealogy model.

Having the distribution $P(T_{2c} = t \mid tree)$ it is also possible to compute its expected value $E(T_{2c} \mid tree)$ denoted as $T_{2c\_agv}|tree$. Additionally, it is possible to trace back lineages of the whole population to the MRCA and therefore to obtain $T_{MRCA}|tree$, as well as the ratios $(T_{2c\_avg} / T_{MRCA})|tree$ and $(T_{MRCA} / T)|tree$. Finally, by simulating many branching processes and averaging over trees generated, let us obtain the corresponding unconditional characteristics H_{T2c}, $P(T_{2c} = t)$ and its expectation $E(T_{2c})$, $P(T_{2c\_agv} = t)$ with the expectation $E(T_{2c\_agv})$, $P(T_{MRCA} = t)$ with the expected $E(T_{MRCA})$, as well as the histograms and the expectations over genealogies of the ratios T_{2c_agv} / T_{MRCA}, and T_{MRCA} / T.

It is also worth to notice that the expectation $E(T_{2c\_agv} / T_{MRCA})$ obtained in the procedure described above, can be used in this model in the equation (2) instead of $E(T_{2c\_agv} / T_{MRCA})$ what will yield a smaller variance estimator. It is justified from the genetic point of view by a clear association of the expectation $E(T_{2c\_agv})$, scaled in (1) by the divergence rate $\delta = \mu / \lambda$,

with the average pairwise mutation difference in a sample d_{avg}. Note also that the simulations which became extinct were excluded from computations, since problems similar to those of dating MRCA of modern humans, are posed in general conditionally on non-extinction – the exception to this rule will be discussed in the section 5.4, where the interest is in the extinct due to genetic drift hypothetical mtDNA of Neanderthals in the modern human gene pool.

The software operating in the second mode stores only the course of population size in the evolution described by a branching process. This mode is used for numerical computation of the distribution of a pair in the Wright-Fisher (3.5:2) or the coalescent (3.5:36) models conditional on N_t. Equation (3.5:2) can be applied directly if the history of N_t is available, whereas in the continuous coalescent model it is possible to apply the Monte Carlo approach by generation of the coalescence times from the distribution conditional on N_t (3.5.34) and repetition of the procedure up to 10^4 times for one simulated branching process.

The conditional histogram which can be obtained in this way is used as the approximation of the conditional distribution $P\,(T_{2c} = t \mid N_t, CM)$, where CM denotes the coalescent model. As in the case of the full genealogy models, the unconditional (with respect to N_t, but obviously conditional with respect to the model used) distributions $P\,(T_{2c} = t \mid CM)$ and $P\,(T_{2c} = t \mid W\text{-}F)$ with $W\text{-}F$ denoting the Wright-Fisher model, are obtained by averaging over many realizations of N_t.

Since in the first mode of operation it is necessary to simulate the population evolution trees which are the dynamic data structures of extremely huge size, predictable only in statistical fashion, and additionally it is required to iteratively generate them several thousand times, the proper administration of the computer operating memory, as well as the time efficiency of the algorithm were the two relevant problems the author had to face in the design of the software. This excluded the use of interpreted languages like Matlab, and even byte-code languages like Java or C#. The author had at his disposal ObjectPascal and C++, and the first of them has been chosen in Borland environment - Delphi.

However, the use of the programming language compiling high-level commands to the native code of the processor, has the drawback of insufficient quality of the built-in random number generator, and this is the third, apart from the memory administration and the time efficiency, problem the author had to overcome. To perform the required number of simulations in an uncorrelated and aperiodic way, there was implemented a generator (Wieczorkowski and Zieliński 1997, Marsaglia *et al*. 1990) being a union of a Fibbonacci generator with period 2^{120} and auxiliary generator with period 2^{24}-1. The aperiodicity length 2^{144} of the resulting generator which additionally fulfills the requirements of all known statistical tests, in particular the tests based on overlapping pairs sparse occupancy (OPSO) method (Marsaglia 1993) was considered as more than satisfactory.

The sufficiency of mentioned generator is based on the fact that for, say 10^5 simulated branching processes used for computing some distribution, and 10^4 generations (for human generation length being approximately 20 years it is equivalent to 200,000 years, covering time comparable to that elapsed from the mtEve, until present) ultimately having not more than 10^{11} individuals (the last number is taken with a margin for simulation of branching process with initial positive fluctuation of the population size which is "frozen" and later exponentially growing to the size much exceeding that predicted by the expected value; certainly such large-scale simulations are feasible only in the second mode) we can expect considerably less than $10^{5+4+11} = 10^{20} < 2^{70}$ invokes of the random number generator. Each such call generates a random number from the uniform (0,1) distribution, transformed to a number from the desired distribution with pgfs given by (4), (5), or (6) and denoting the number of progeny of given individual. Moreover, the requirement of the generator for the representation of at least 16 bit integers and at least 24 bit mantissas for variable-precision numbers is always satisfied in contemporary computers, and it guarantees the invariance of the generation with respect to details of computer's representation of numerical values.

The formal comparison of experimental cumulative distributions $F_{sim}(t)$ with the theoretical O'Connell cumulative distribution $F_{OC}(t)$ is performed using the Kolmogorov-Smirnov test with statistics

$$D_1 = \max_t \left| F_{sim}(t) - F_{OC}(t) \right| \tag{5.3:11}$$

with null hypothesis H_0 stating that $F_{sim}(t)$, obtained from n non-extinct simulations of branching process is equal to $F_{OC}(t)$. Similar tests were conducted for equality of two experimental cumulative distributions $F_{sim1}(t)$ and $F_{sim2}(t)$ based on numbers of non-extinct simulations n_1 and n_2, respectively. Then the test statistic is given by

$$D_2 = \max_t \left| F_{sim1}(t) - F_{sim2}(t) \right|, \tag{5.3:12}$$

i.e. it has similar form, however there are different critical values for D_1 and D_2.

Assuming that numbers n, n_1, $n_2 > 40$, the critical value for D_1 is

$$D_{1\_\alpha} = c(\alpha)\frac{1}{\sqrt{n}} \tag{5.3:13}$$

and the critical value for D_2 is

$$D_{2\_\alpha} = c(\alpha)\sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \tag{5.3:14}$$

respectively. In above equations α denotes the significance level of the test, and $c(\alpha)$ is given in the Table 1.

To obtain the estimates of the time to MRCA from the models discussed there are considered the average pairwise mutation differences d_{avg} and the genetic divergence rate δ

computed from a sample of 663 mtDNA sequences of modern humans and their homologs sequenced from the Neanderthal fossils (Krings *et al*. 1999). These sequences were taken from the hypervariable control region I (HVRI) and the hypervariable control region II (HVRII) of the mtDNA, respectively. After elimination of insertions and deletions the concatenated sequences yielded 600bp in total, as reported by Krings *et al*. (1999). In this sample the average pairwise number of the segregating sites is equal to 35.3 ± 2.3. Therefore the average genetic distance is equal to $d_{avgM\text{-}N} = 5.9$ %.

The divergence in contemporary humans results in an average number of segregating sites equal to 10.9 ± 5.1 and thus the average mutation difference among contemporary humans is equal to $d_{avg} = 1.8$ %. For comparison there is also presented the average mutation difference among modern humans calculated originally by O'Connell (1995) to be equal 2.8 %, but it is not considered further, because of a much smaller sample size of 19 humans used by O'Connell.

Table 5.3:1
Parameters $c(\alpha)$ used for computing the critical values of the Kolmogorov-Smirnov test

	Confidence level		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$c(\alpha)$	1.22	1.36	1.63

The average mutation difference between *H. neanderthalensis* and *H. sapiens*, about 3 times greater than that among contemporary humans, is still small enough to allow ignoring reverse mutations occurring in both lineages from the time of their divergence T_d some 370,000 years ago (Noonan *et al*. 2006).

Therefore, by applying the infinite sites model, it is possible to compute the rate of divergence as $\delta = d_{avgM\text{-}N} / T_d \approx 0.059/370{,}000 = 1.6\times10^{-7}$ mutations per nucleotide per year. This estimate is slightly above the upper bound of 95 % confidence interval $[5.9\times10^{-8}, 1.4\times10^{-7}]$ reported by Adachi and Hasegawa (1995), indicating that recent discoveries based on the results of the Neanderthal Genome Project suggest faster molecular clock. This project yielded some successful sequencings of the nuclear DNA of the Neanderthals having equivocal interpretations in terms of the Neanderthal admixture in the modern human gene pool (Plagnol and Wall 2006, Pennisi 2006, Noonan *et al*. 2006, Pennisi 2007), however since these sequences were subject to the recombination, they are not considered in greater detail in the study based on the branching processes genealogy.

For consistency of the comparison, the results of the experiments performed are presented in the discrete reversed time expressed in generation units. The results of the models which

traditionally use differently measured time are scaled before presentation to satisfy this common unifying requirement. Note also that despite the discrete nature of the time, the distributions are drawn in the form of continuous curves because such artificially introduced continuity visually helps to trace any particular distribution, separating it from the others presented in the same plot.

Let us start with the illustration of the fact that the model with full genealogy yields visually undistinguishable distributions $P\,(T_{2c} = t)$ (see Fig.1), $P\,(T_{2c\_avg} = t)$ (see Fig.2), $P\,(T_{MRCA} = t)$ (see Fig.3) and $P\,(T_{2c\_avg} / T_{MRCA} = x)$ (see Fig.4) regardless of the offspring distribution, and thus its variance, for the same mean number of progeny.

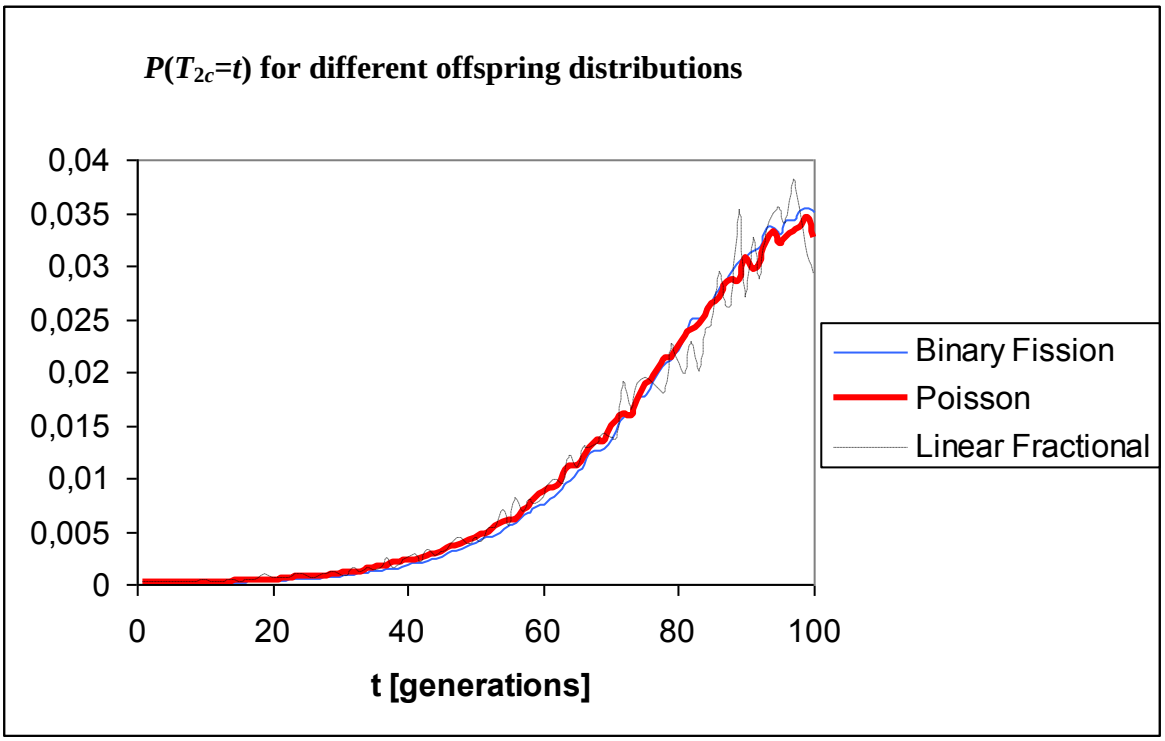


Fig. 5.3:1. Distributions of T_{2c} computed in the full genealogy model
Rys. 5.3:1. Rokłady T_{2c} obliczone w modelu pełnej genealogii

Interestingly, this visual identity remains true in spite of equivocal results (see Table 2) of the Kolmogorov-Smirnov test for pairwise comparison of cumulative distributions $F_{sim}\,(t) = P\,(T_{2c} < t)$ obtained using the offspring distributions with pgfs given by (4), (5), or (6).

Table 2 presents results computed in the full genealogy model of branching processes with different offspring distributions, serving as headers of rows. Bold font is used to indicate the critical values D_{2_α} exceeding the value of the corresponding statistic D_2. The comparison of the shapes of all these distributions and the deterministic distribution $P\,[T_{2c} = t \mid E\,(N_t)]$ for the Poisson offspring distribution is given in the Fig. 5.

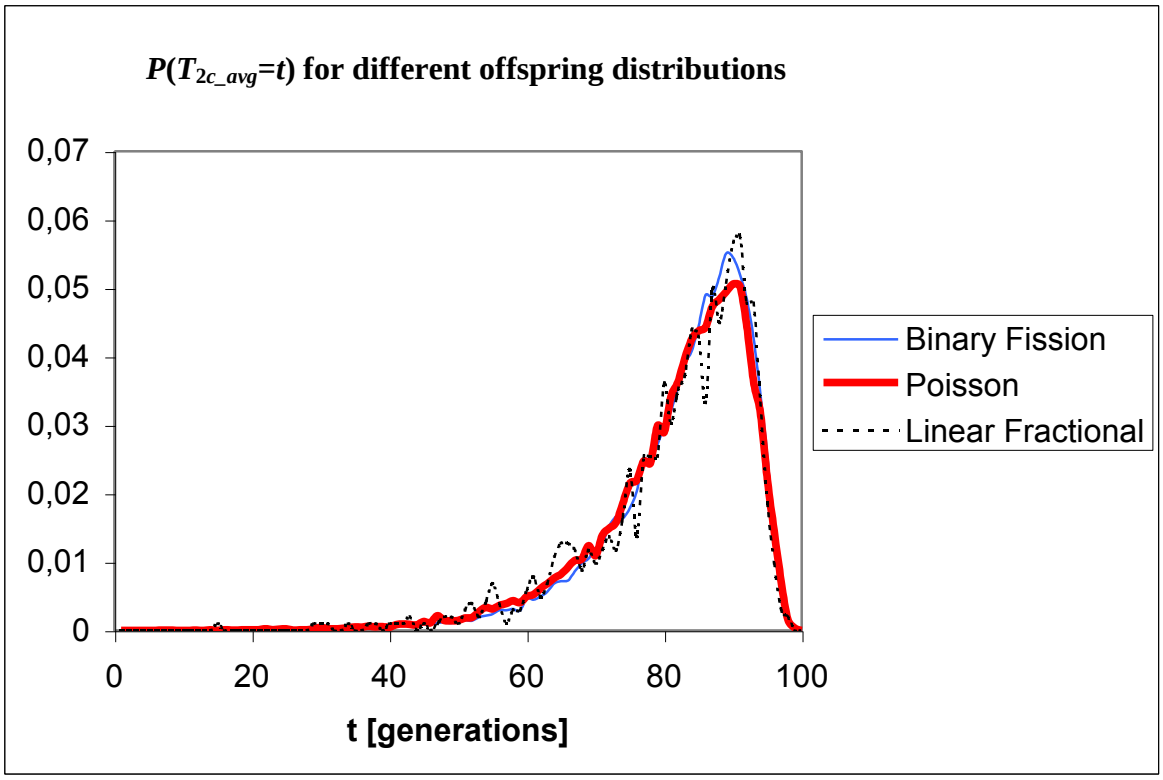


Fig. 5.3:2. Distributions of T_{2c_avg} computed in the full genealogy model
Rys. 5.3:2. Rokłady T_{2c_avg} obliczone w modelu pełnej genealogii

Table 5.3:2
Results of Kolmogorv-Smirnov test for a pairwise comparison of the cumulative distributions F_{sim1} and F_{sim2} of T_{2c}

F_{sim1}	F_{sim2}	n_1	n_2	D_2	$D_{2\_\alpha = 0.1}$	$D_{2\_\alpha = 0.05}$	$D_{2\_\alpha = 0.01}$
BF	*P*	33164	17766	0.0224	0.0113	0.0126	0.0152
BF	*LF*	33164	1024	0.0295	**0.0387**	**0.0432**	**0.0517**
P	*LF*	17766	1024	0.0111	**0.0392**	**0.0437**	**0.0524**

More importantly, (see Fig. 6, 7, and 8) the distributions $P(T_{2c} = t)$ obtained for any offspring distribution are also visually identical to the O'Connell limiting distribution for as little as 100 generations when the O'Connell parameter $\alpha = 10$. Although the choice of $\alpha = 10$ seems somewhat arbitrary, the analysis of the O'Connell (1995) proved than any value between 10 and 14 is feasible and has little effect on the estimates. Therefore, for simplicity, and for clearer demonstration of the stochastic effects there was chosen value of $\alpha = 10$, yielding branching processes closer to the critical as compared to those with greater (but not

exceeding 14 according to the O'Connell's feasibility analysis results) values of this parameter.

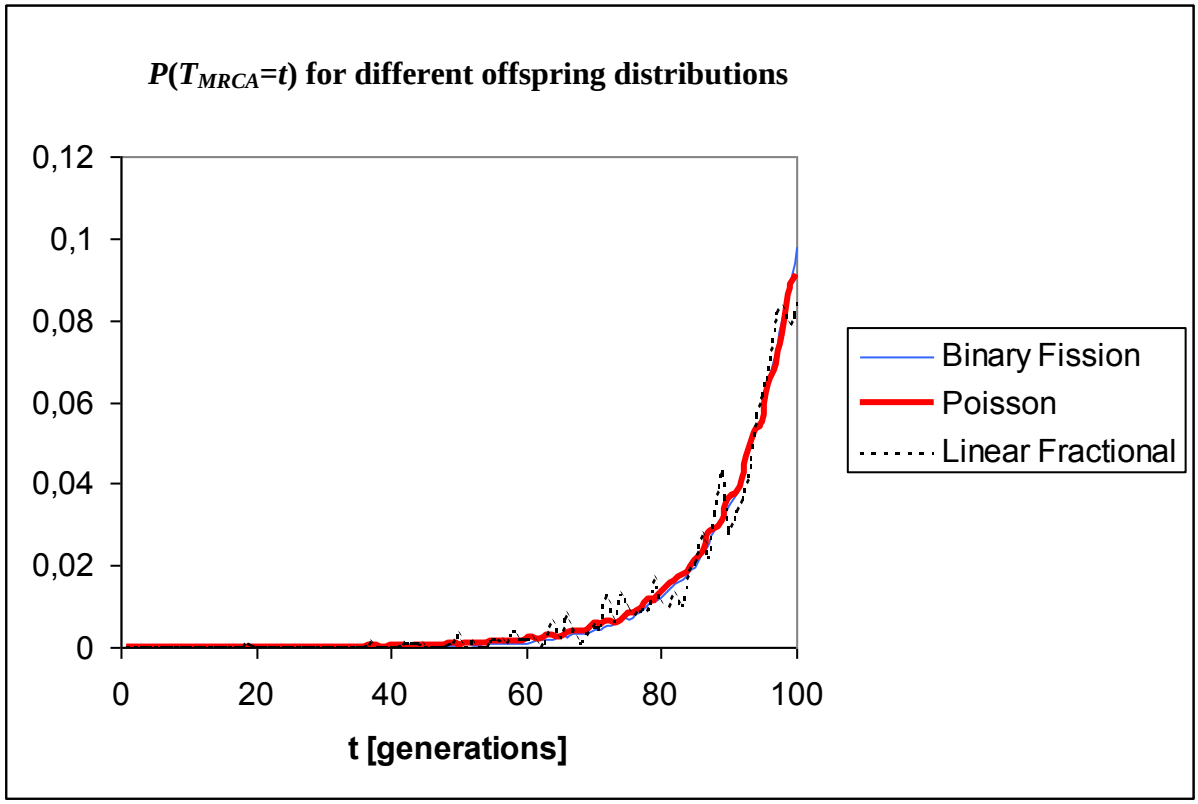


Fig. 5.3:3. Distributions of T_{MRCA} computed in the full genealogy model
Rys. 5.3:3. Rozkłady T_{MRCA} obliczone w modelu pełnej genealogii

Despite equivocal results of the Kolmogorov-Smirnov tests (see Table 3, where bold font is used to indicate the critical values D_{1_α} exceeding the value of the corresponding statistic D_1), the visual inspection of Fig. 6, 7, and 8, together with comparison of the expectations presented in Table 4 ensures that the limiting O'Connell distribution $P\,(T_{2c} = t \mid OC)$ almost perfectly mimics the distributions of T_{2c} obtained in the full genealogy model for 100 generations. Therefore it is possible to map the expectation of T_{MRCA} available directly only in the full genealogy model on a time scale of the O'Connell model and therefore it is possible to compute the expectation of the ratio T_{MRCA} / T required in (3) not only in the full genealogy model, but also in the O'Connell model.

Because of the asymptotic character of the above results, they remain valid for arbitrary number of generations exceeding 100 for which the validity of asymptotic predictions was experimentally verified (if only the parameters used in the O'Connell model remain the same). Therefore, even if it is not possible to compute it directly in the full genealogy model, by indirect combining with the limiting O'Connell results it is possible to obtain the ratio

T_{MRCA} / T also for the number of generations of the order 10^4, corresponding to the time elapsed from the death of the mtEve until present.

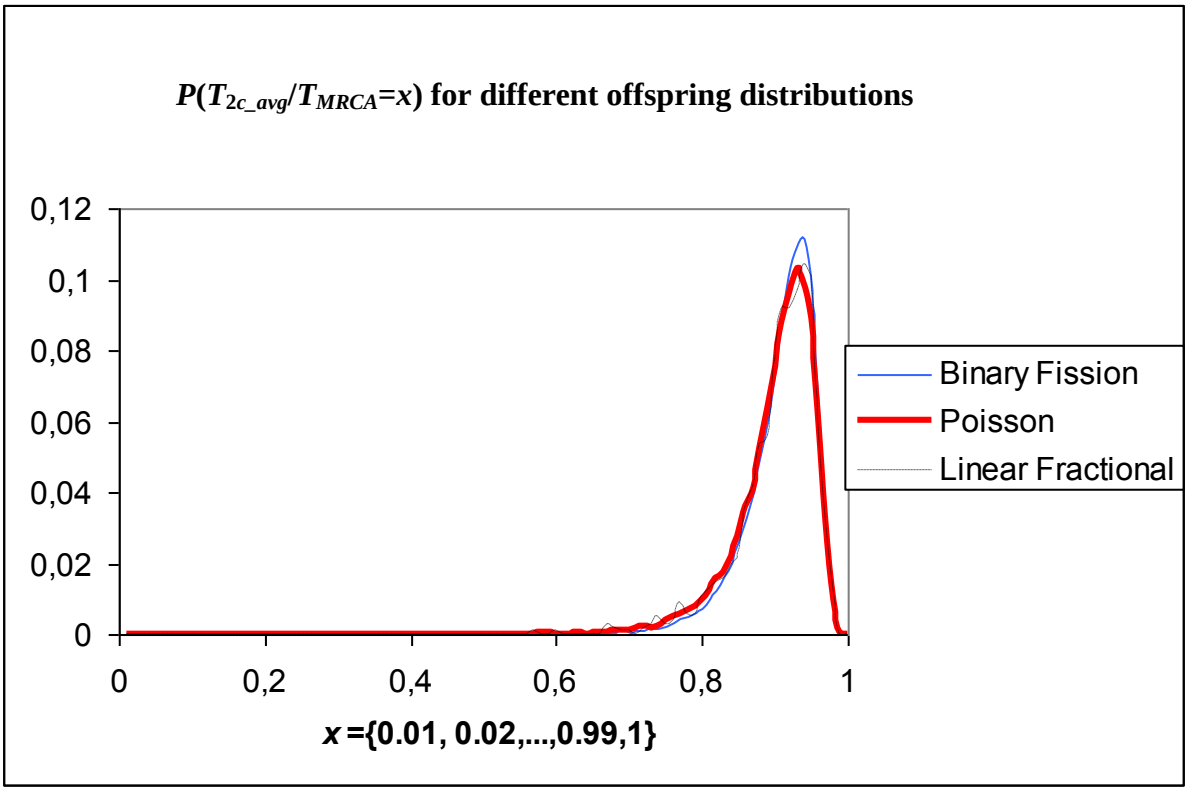


Fig. 5.3:4. Distributions of the ratio T_{2c_avg} / T_{MRCA} computed in the full genealogy model
Rys. 5.3:4. Rozkłady stosunku T_{2c_avg} / T_{MRCA} obliczone w modelu pełnej genealogii

Table 5.3:3
Results of the Kolmogorv-Smirnov test for T_{2c} distributions F_{sim} computed in the full genealogy model of branching processes with different offspring distributions compared to the limiting O'Connell distribution $F_{theoretical}$

F_{sim}	$F_{theoretical}$	n	D_1	$D_{1\_\alpha = 0.1}$	$D_{1\_\alpha = 0.05}$	$D_{1\_\alpha = 0.01}$
BF	*OC*	33164	0.0120	0.0067	0.0075	0.0090
P	*OC*	17766	0.0118	0.0092	0.0102	**0.0122**
LF	*OC*	1024	0.0187	**0.0381**	**0.0425**	**0.0509**

In this work, there was also studied the relationship of the Wright Fisher discrete model with the continuous coalescent model applied to stochastic population histories approximated by slightly supercritical branching process. The corresponding distributions of the time to coalescence of two individuals T_{2c} for binary fission offspring distribution are presented in

Fig. 9 together with superimposed O'Connell distribution. Similarly, a comparison of the coalescence distributions dependent on the model used are presented in Fig. 10 for Poisson offspring distribution, and in Fig. 11 for the linear fractional offspring distribution.

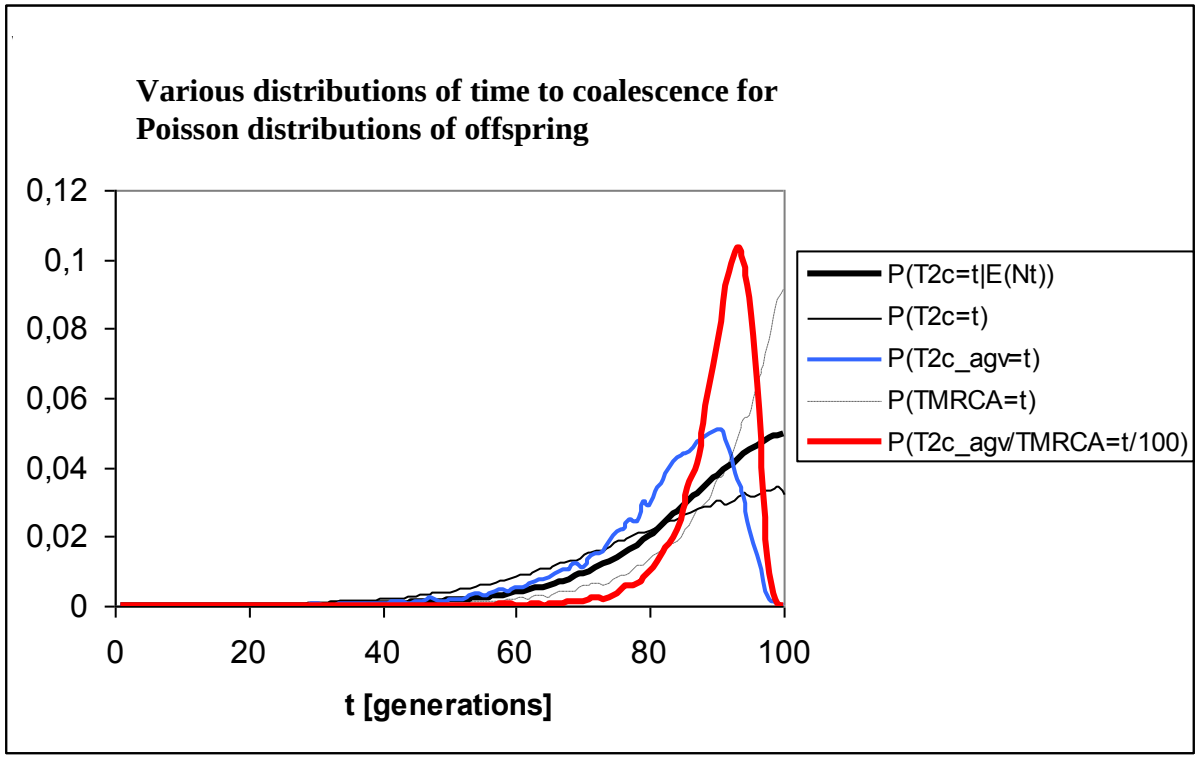


Fig. 5.3:5. General comparison of the coalescence distributions obtained in the full genealogy model for the Poisson offspring distribution
Rys. 5.3:5. Ogólne porównanie rozkładów koalescencji otrzymanych w modelu pełnej genealogii dla Poissonowskiego rozkładu potomstwa

Table 5.3:4
Expectations of the ratio $T_{2c} / T \pm SD$ in the O'Connell and the full genealogy models

Model	$E\ (T_{2c} / T)$
O'Connell	0.8054 ± 0.1591
Full genealogy with BF progeny	0.8097 ± 0.1585
Full genealogy with P progeny	0.8008 ± 0.1645
Full genealogy with LF progeny	0.8002 ± 0.1662

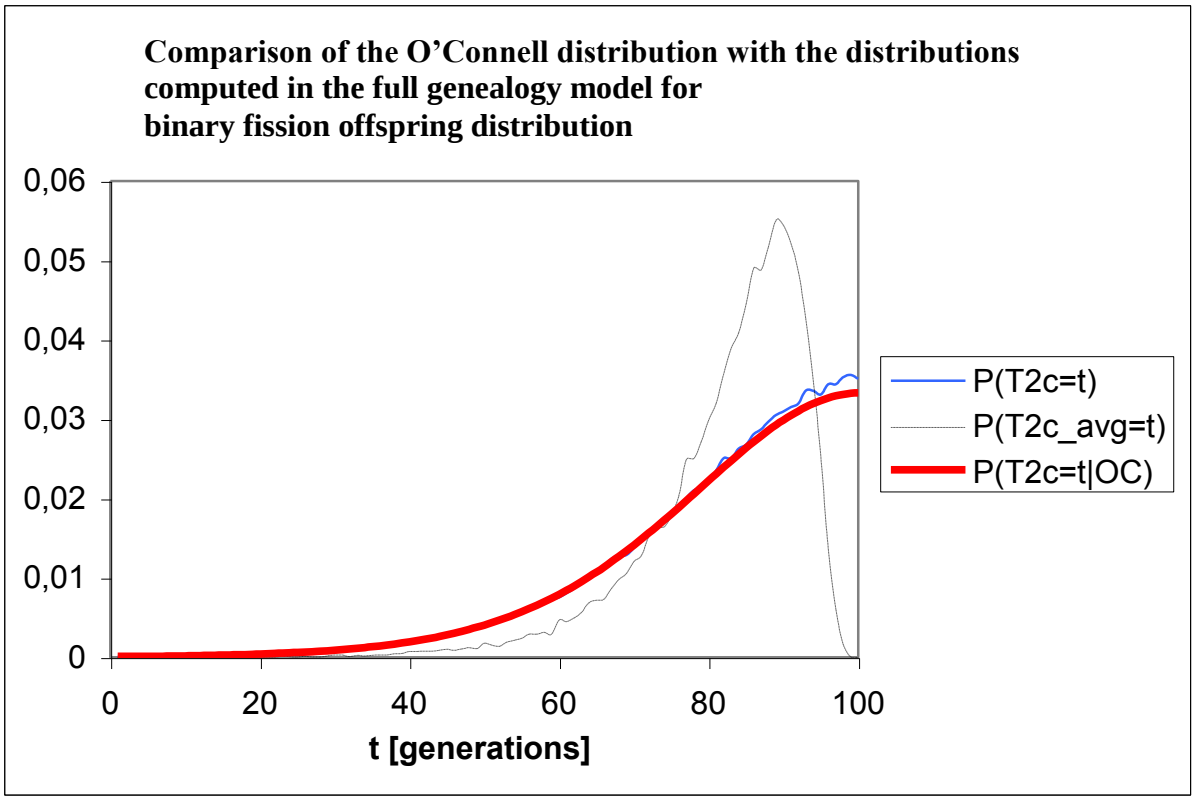


Fig. 5.3:6. Comparison of the distributions of T_{2c} in the full genealogy model and in the limiting O'Connell model for BF offspring distribution
Rys. 5.3:6. Porównanie rozkładów T_{2c} w modelu pełnej genealogii oraz w granicznym modelu O'Connella dla rozkładu potomstwa BF

The inspection of Fig. 9 - 11 reveals that both models considered deviate from the O'Connell model for offspring distributions other than Poisson. Since the continuous coalescent model is equivalent with the diffusion process limit, which in turn is dependent on the variance of progeny, this result can be easily explained by the variances of binary fission and linear fractional distributions deviating from the variance of Poisson distribution in opposite directions. There is one more interesting fact which can be observed. Namely, for times t close to T (corresponding to the beginning of branching process) the continuous approximation assumed in the coalescent theory lacks its validity and the distribution differs more and more from the Wright-Fisher distribution. This is finally reflected in the atom of probability at $t = T$ required for probabilities to sum to one. However, despite this visually striking feature and the Kolmogorov-Smirnov test results, clearly differentiating between the distributions (Table 5), the expectations of $T_{2c}|WF$ and $T_{2c}|CM$ remain very similar (see Table 6). For completeness of the study there is also presented in Fig. 12 a comparison of the deterministic distributions $P\,(T_{2c}\,|\,E\,(N_t))$ for different offspring sampling schemes, together with the limit O'Connell distribution.

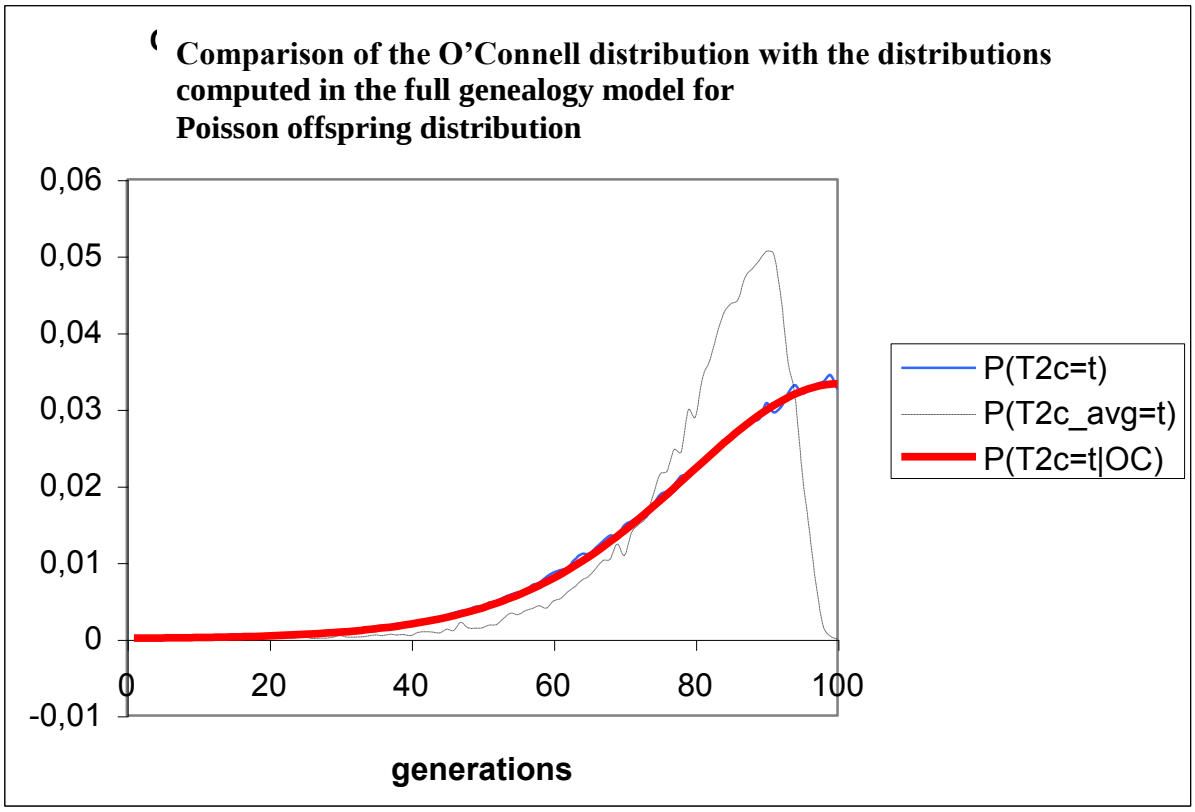


Fig. 5.3:7. Comparison of the distributions of T_{2c} in the full genealogy model and in the limiting O'Connell model for Poisson offspring distribution
Rys. 5.3:7. Porównanie rozkładów T_{2c} w modelu pełnej genealogii oraz w granicznym modelu O'Connella dla Poissonowskiego rozkładu potomstwa

Table 5.3:5
Results of the Kolmogorov-Smirnov test for comparison of the cumulative distribution F_{sim1} computed in the Wright-Fisher model and F_{sim2} computed in the coalescent model with different offspring distributions, serving as headers of rows

F_{sim1}	F_{sim2}	n_1	n_2	D_2	$D_{2\_\alpha = 0.1}$	$D_{2\_\alpha = 0.05}$	$D_{2\_\alpha = 0.01}$
BF|WF	*BF|CM*	33195	33126	0.025	0.009	0.010	0.013
P| WF	*P|CM*	17342	17520	0.035	0.013	0.015	0.017
LF| WF	*LF|CM*	9916	9922	0.071	0.017	0.019	0.023

The deterministic population growth used for generation of the distributions presented in Fig. 12 is modeled by taking the expectation of population sizes resulting from the realizations of branching processes with different offspring distributions. After performing comparisons between models, let us focus on the results of the full genealogy model and let

us present the expectations and their standard deviations of the times directly available only in this latter model. These results are given in Table 7.

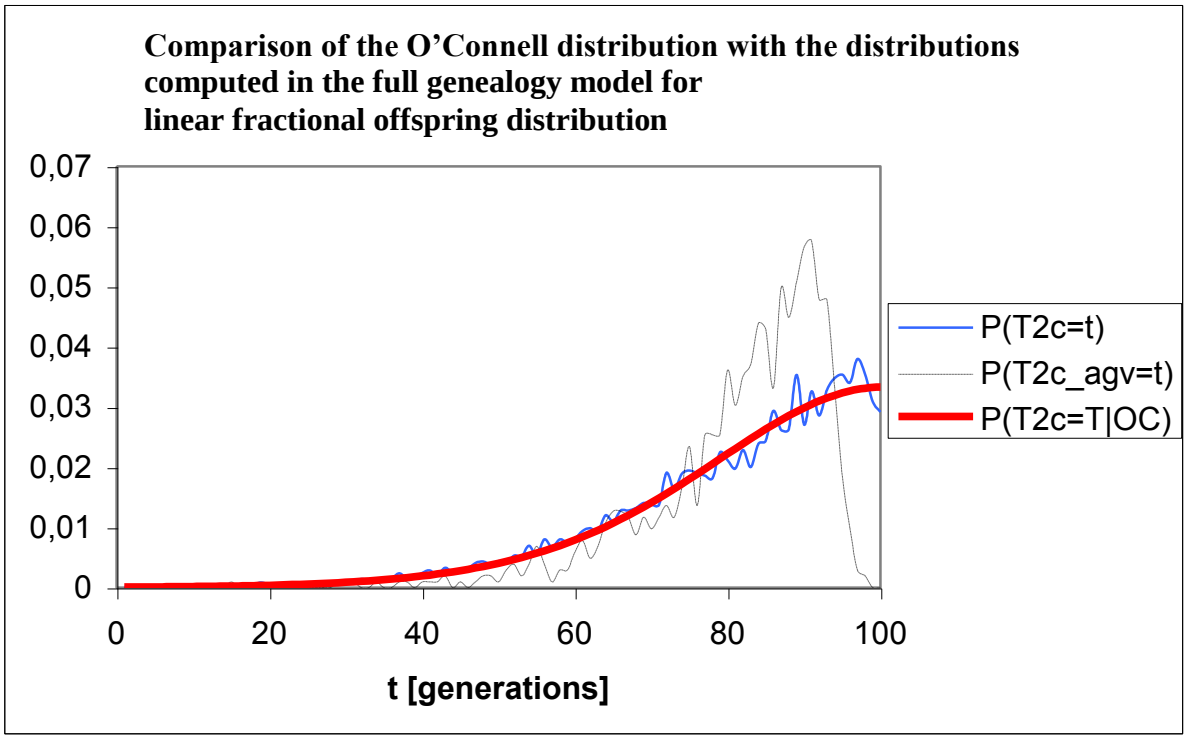


Fig. 5.3:8. Comparison of the distributions of T_{2c} in the full genealogy model and in the limiting O'Connell model for LF offspring distribution
Rys. 5.3:8. Porównanie rozkładów T_{2c} w modelu pełnej genealogii oraz w granicznym modelu O'Connella dla rozkładu potomstwa LF

Table 5.3:6
Comparison of the expectations of T_{2c} / T computed in the Wright-Fisher and the coalescent models for different offspring distributions

Progeny distribution	$E\,(T_{2c}/T \mid WF)$	$E\,(T_{2c}/T \mid CM)$
BF	0.7497	0.7585
P	0.8005	0.8078
LF	0.8454	0.8550

Using the O'Connell model with T_{MRCA} moment mapped according to the full genealogy model, it is possible to estimate the time to the mtEve. The estimates of this time, assuming $\delta = 1.6 \times 10^{-7}$ and $d_{avg} = 0.018$, for different population histories, are given in Table 8. It is

visible that the time is of the order of 10^4 generations. The simulated distributions in the Wright-Fisher model with different offspring distributions compared to the O'Connell distribution for this number of generations are presented in Fig. 13. Fig. 14 presents distributions for similar time span, however it shows the influence on the reproduction success of the environment variable in time. The inhomogeneity in time was introduced by changing the expected number of offspring with parameters $\sigma_{1e} = 0.09 \times \mu$ and $\sigma_{2e} = 3 \times \sigma_{1e} = 0.27 \times \mu$, where μ is the expected number of progeny.

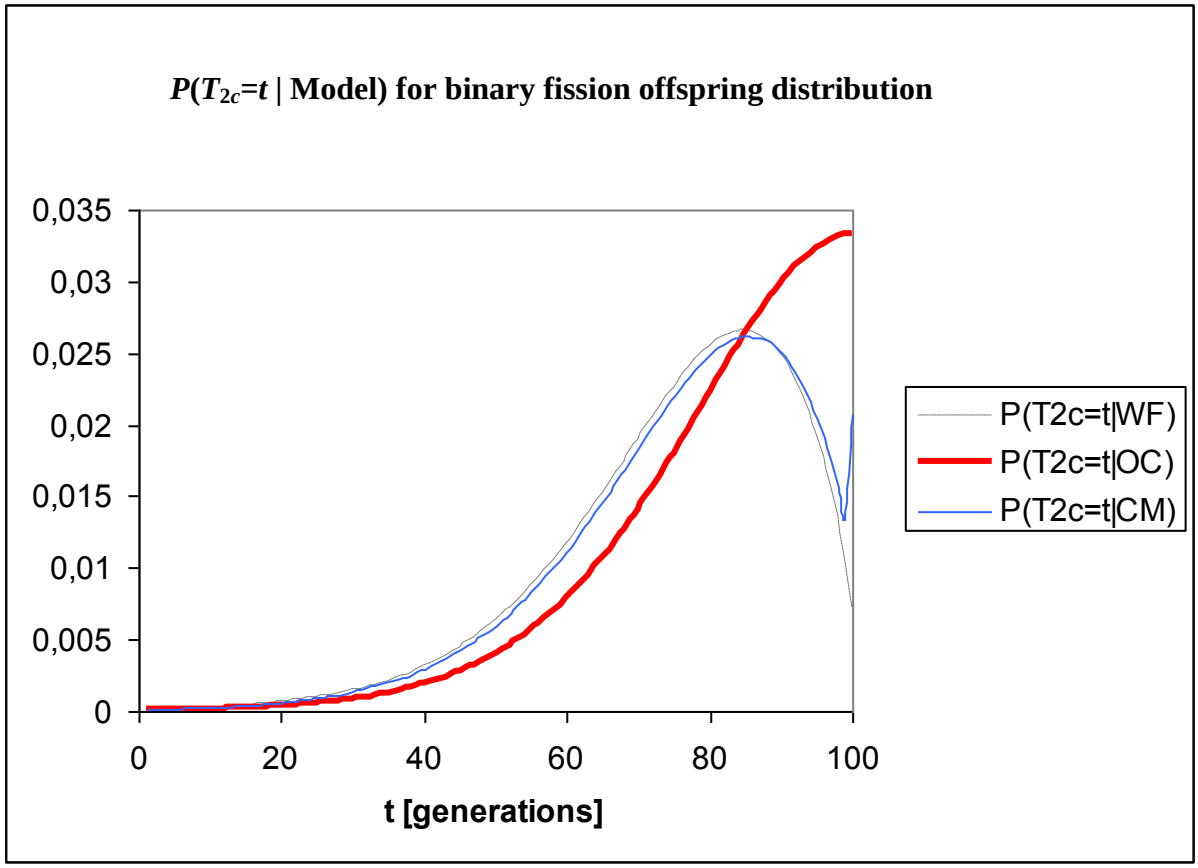


Fig. 5.3:9. Comparison of distributions of T_{2c} computed in the Wright-Fisher, the coalescent and the O'Connell models for BF offspring distribution
Rys. 5.3:9. Porównanie rozkładów T_{2c} obliczonego w modelu Wrighta-Fishera, koalescentu oraz O'Connella dla rozkładu potomstwa BF

These results contribute to the conclusion that random environmental changes have influence on the coalescence time distribution similar to that caused (somewhat surprisingly) by a decrease of the variance of offspring distribution compared to the Poisson offspring distribution, however the influence is spread over a longer time. This larger span of the influence is observed because the environmental stochasticity, contrary to the demographic stochasticity, is not eliminated by the increase of the size of population.

Table 5.3:7

Expectations of different ratios of the coalescence times and their standard deviations computed in the full genealogy model for various distributions of progeny

Parameter	*BF*	*P*	*LF*
$E\ (T_{2c} / T)$	0.8097 ± 0.1585	0.8008 ± 0.1645	0.8002 ± 0.1662
$E\ (T_{2c\_avg} / T)$	0.8097 ± 0.1057	0.8009 ± 0.1124	0.8001 ± 0.1150
$E\ (T_{MRCA} / T)$	0.9094 ± 0.0950	0.9032 ± 0.1011	0.9017 ± 0.1040
$E\ (T_{2c\_avg} / T_{MRCA})$	0.9068 ± 0.0482	0.9027 ± 0.0532	0.9035 ± 0.0535

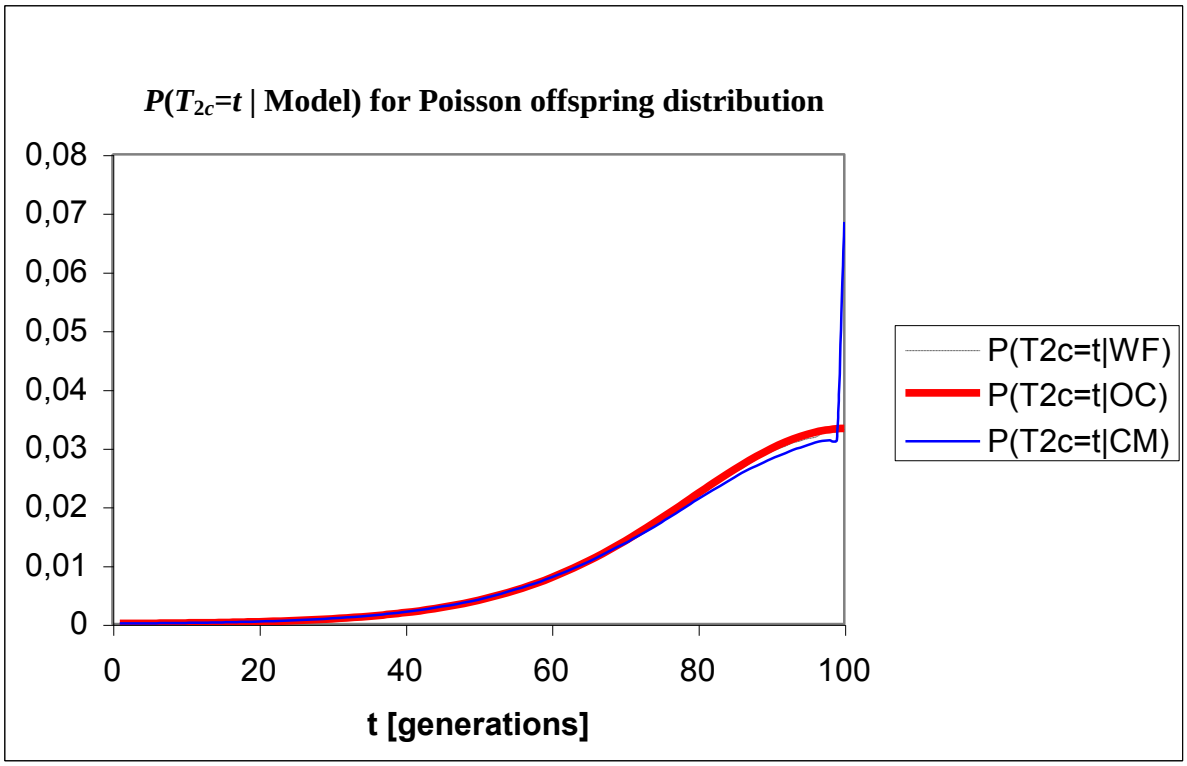


Fig. 5.3:10. Comparison of distributions of T_{2c} computed in the Wright-Fisher, the coalescent and the O'Connell models for Poisson offspring distribution

Rys. 5.3:10. Porównanie rozkładów T_{2c} obliczonego w modelu Wrighta-Fishera, koalescentu oraz O'Connella dla Poissonowskiego rozkładu potomstwa

Comparison of the numbers in Table 8 with the expectation 163×10^3 years and the corresponding 95 % confidence interval $[111 \times 10^3, 260 \times 10^3]$ reported by Krings *et al.* (1999) shows that all stochastic model predictions fall into phylogenetically obtained interval,

although particular coalescence time distributions vary among models considered. Moreover, because of faster molecular clock used in this study (caused by the expectation of the modern humans and Neanderthals split shifted towards present), all expectations are considerable more recent than those of Krings *et al*. (1999).

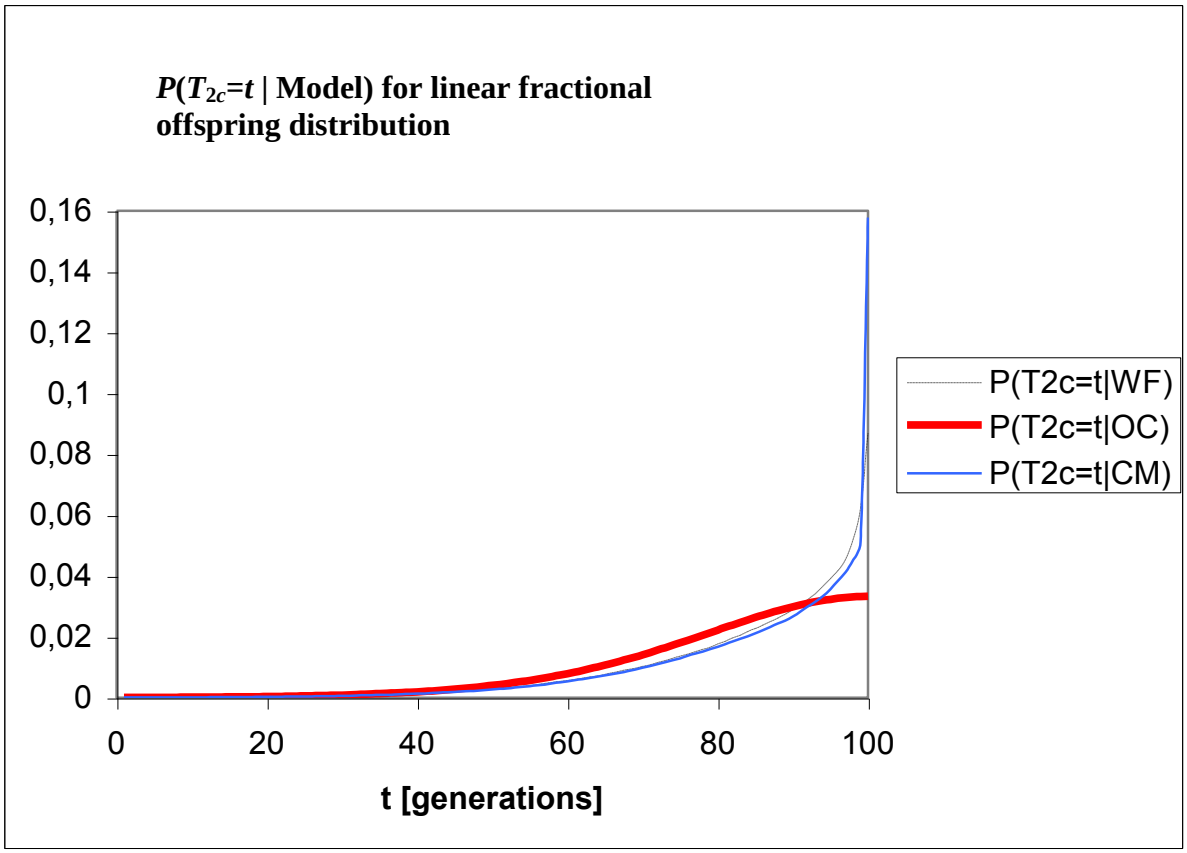


Fig. 5.3:11. Comparison of distributions of T_{2c} computed in the Wright-Fisher, the coalescent and the O'Connell models for LF offspring distribution
Rys. 5.3:11. Porównanie rozkładów T_{2c} obliczonego w modelu Wrighta-Fishera, koalescentu oraz O'Connella dla rozkładu potomstwa LF

Table 9 presents the data required to compute the expectation of T_{MRCA} and the 95 % confidence interval in the full genealogy model with the use of equation (2). In Table 9, the column *value*$_-$ gives such bound for the given parameter, which yields the lower bound for T_{MRCA_y}. Correspondingly, the column *value*$_+$ gives such bounds, which yield the upper bound for T_{MRCA_y}. For d_{avg} and d_{HN}, the normal distribution is approximately assumed (see Krings *et al*. 1997), and therefore the 95 % interval bounds are computed according to $2 \times \sigma$ rule, where σ denotes the standard deviation of the corresponding distribution.

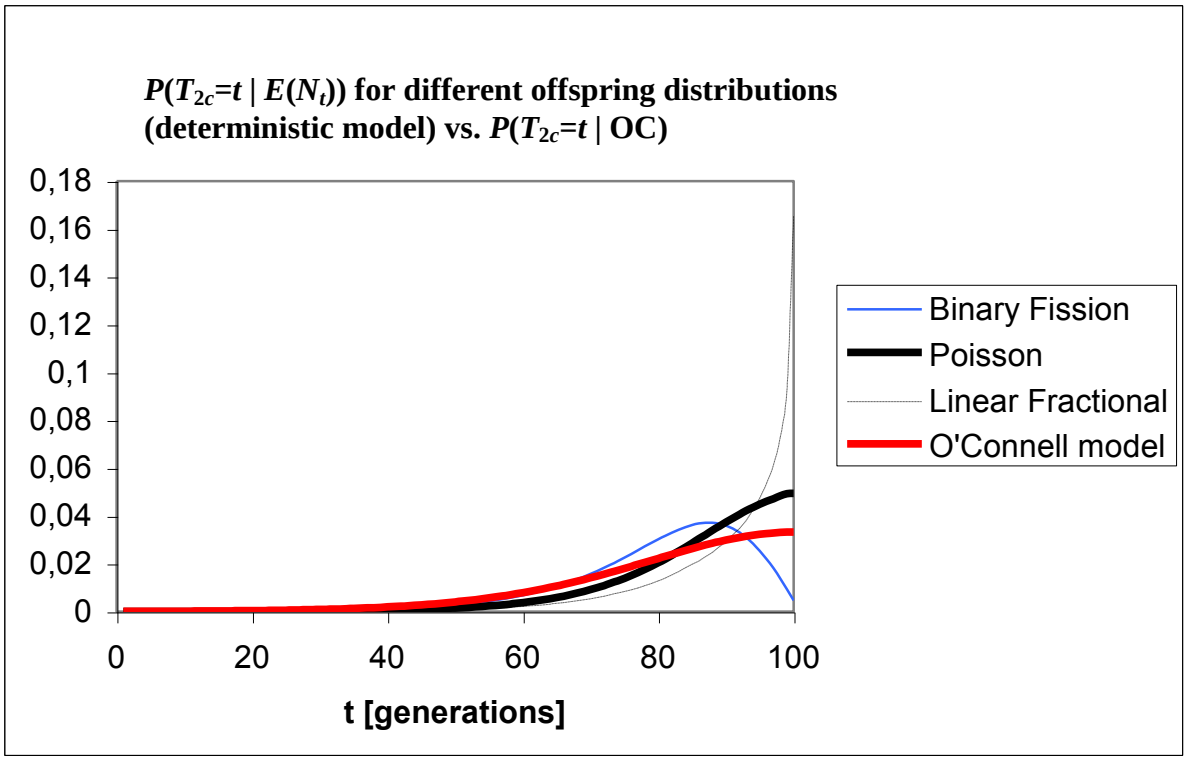


Fig. 5.3:12. Comparison of distributions of T_{2c} computed in the Wright-Fisher model for deterministic population growth
Rys. 5.3:12. Porównanie rozkładów T_{2c} obliczonego w modelu Wrighta-Fishera dla deterministycznego wzrostu populacji

Table 5.3:8
Expectations of the time to MRCA of modern humans computed in the O'Connell, the full genealogy, the Wright-Fisher and the coalescent models

Model	$\hat{T}_{MRCA\_y}$ [thousands of years]
O'Connell limit	128
Full Genealogy, Binary Fission	126
Full Genealogy, Poisson	126
Full Genealogy, Linear Fractional	126
Wright-Fisher, Binary Fission	137
Wright-Fisher, Poisson	129
Wright-Fisher, Linear Fractional	122
Coalescent, Binary Fission	136
Coalescent, Poisson	127
Coalescent, Linear Fractional	120

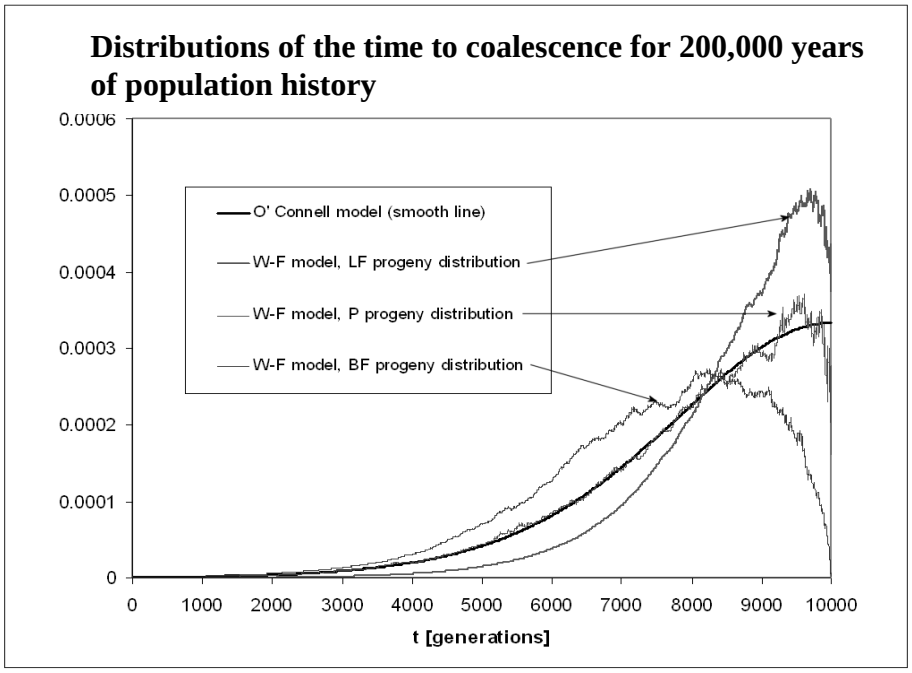


Fig. 5.3:13. Distributions computed in the Wright-Fisher model for stochastic population growth modeled by the branching process encompassing 10^4 generations

Rys. 5.3:13. Rozkłady obliczone w modelu Wrighta-Fishera dla stochastycznych wzrostów populacji modelowanych przez process łańcuchowy z 10^4 pokoleniami

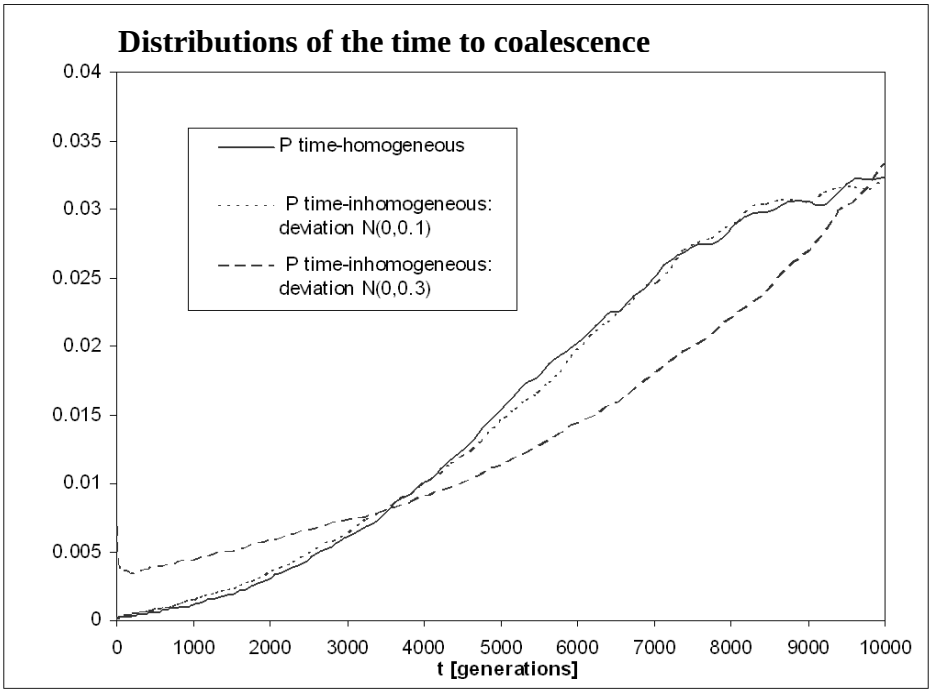


Fig. 5.3:14. Influence on the coalescence distributions of changes in the reproduction success modeled by Poisson distribution with randomly changing mean and thus variance

Rys. 5.3:14. Wpływ na rozkłady koalescencji zmian w sukcesie reprodukcyjnym modelowanym przez Poissonowski rozkład ze zmieniającą się losowo wartością oczekiwaną a zatem i wariancją

Table 5.3:9

Expectation and 95 % confidence interval of T_{MRCA_y}

Parameter	$value_-$	*expectation*	$value_+$
d_{agv}	0.01751	0.01817	0.01883
d_{HN}	0.0669	0.0592	0.0515
$T_{MRCA\_HN\_y}$	200×10^3	370×10^3	600×10^3
δ	3.35×10^{-7}	1.6×10^{-7}	0.86×10^{-7}
T_{2c} / T_{MRCA}	1	0.9	0.6
T_{MRCA_y}	52×10^3	$\mathbf{126 \times 10^3}$	365×10^3

As it can be seen there is obtained the expectation of $T_{MRCA} = 126 \times 10^3$ years with the confidence interval $[52 \times 10^3, 365 \times 10^3]$. The confidence interval is computed in a conservative way, i.e., to compute the lower bound of T_{MRCA} there is used the lower bound of d_{avg} and the upper bounds of δ and T_{2c} / T_{MRCA}, whereas for the upper bound of T_{MRCA} there is used the upper bound of d_{avg} and the lower bounds of δ, and T_{2c} / T_{MRCA}, respectively.

Additionally to compute the lower bound of δ there is used the lower bound of the average mutation difference between modern humans and Neanderthals d_{HN} and the date of the split of the two species 200×10^3 years ago, while for the upper bound of δ there is used the upper bound of d_{HN} and the date of the split 600×10^3 years ago. These dates are estimates of the lower and upper bounds of the confidence interval for the modern humans and Neanderthals populations split, as reported by Noonan *et al*. (2006).

5.4. Neanderthal controversy

The coexistence of Neandertals with the Upper Paleolithic anatomically modern humans is a basis for the intriguing problem about the interbreeding between the two (sub)species. This issue is at least as inspiring, as the hypothetical physiognomy of Neandertals - note the change in the reconstruction of Neanderthal face from the earliest, resembling an ape (Fig. 1), to the one of the most recent, resembling the modern human (Fig. 2). Whatever the answer to mentioned problems, some 30,000 years ago, Europe became a scenery of a drama of our closest relatives – after several thousand years of coexistence with *H. sapiens*, the Neandertals had gone. Are their genes still present in the genome of modern humans?

Many mtDNA-based studies, from the earliest (Krings et al. 1997, 1999) to the most recent (Briggs et al. 2009), indicate that *H. Neanderthalensis* is an outgroup in the mtDNA polymorphism of present-day humans. However, after first sequencings of the mtDNA from Neandertal fossils, the resulting phylogenetic tree was erroneously interpreted as an evidence of no interbreeding (for example Krings et al. 1999). While no interbreeding could be the cause for the observed pattern, it cannot be excluded that the mtDNA polymorphism 30 000 years ago was of a different type, and the currently observed pattern is the result of the genetic drift. This latter hypothesis is even more probable in the light of studies based on nuclear DNA in Neandertal genome project. The extent of Neandertal ancestry in modern humans has been estimated by Green et al. (2010) to be between 1 and 4%. The report of Wall et al. (2009), by indicating that the amount of archaic ancestry is about 12%, can suggest that there were also more ancient gene flows (probably from *H. erectus*) to *H. sapiens*.

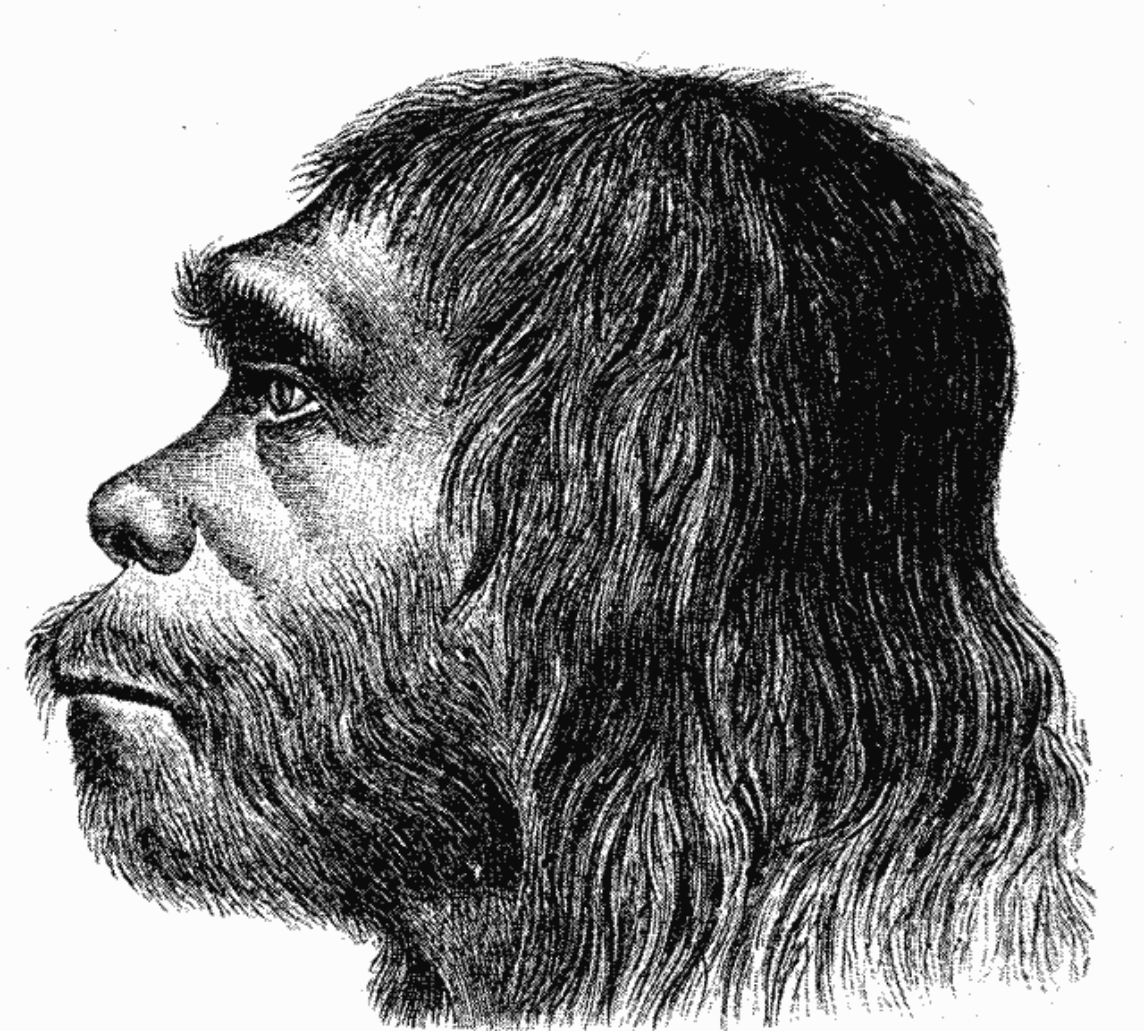


Fig. 5.4:1. The first re construction of Neanderthal. [Picture in public domain]
Rys. 5.4:1. Pierwsza rekonstrukcja Neandertalczyka [Rysunek z *public domain*]

Consider a family of slightly supercritical time-homogeneous Markov branching processes in which the expected numbers of offspring per individual is equal to $E(\xi_0) = 1 + \alpha/T + o(1/T)$ and the corresponding variance is equal to $Var(\xi_0) = \sigma^2 + O(1/T)$, as $T \to \infty$. Such branching process (see Fig. 3.6:1) represents the evolution of Neanderthal mtDNA within the post-Neanderthal modern human population (Fig. 3) after hypothetical admixture. The branching process modeling the Neanderthal mtDNA within the modern human population can become extinct because of the genetic drift. This can happen even with the supercritical process, for which the extinction is not sure, but on the other hand, it is not impossible.

Fig. 5.4:2. Recent re construction of the Neanderthal child. [Picture in public domain]
Rys. 5.4:2. Najmłodsza rekonstrukcja Neandertalczyka [Rysunek z *public domain*]

Let us denote the number of individuals in the process at time t by Z_t. As t we consider the time 30,000 years ago, when the Neanderthals disappeared and their hypothetical admixture in a gene pool of modern humans was a subject to the genetic drift, with no further Neandertal contribution.

Assume the duration T of the branching process to be 200,000 years. Such process is modeling the evolution of *H. sapiens* mtDNA from the MRCA (mtEve) dated to live around 175 000 years ago, with $T_2 = 150\ 000$ years, where T_2 denotes the time to coalescence of a pair of randomly picked mtDNA from a sample of contemporary modern humans. These values are assumed based on results of Cyran and Kimmel (2010), provided that the time to the most recent ancestor of modern humans and Neandertals is 511 000 years ago (Briggs et al. 2009). The mtDNA data used for the inference was taken from Green et al. (2008). Then, the times t and T expressed in the number of generations are: $t = 1\ 500$ generations, $T = 10\ 000$ generations, respectively.

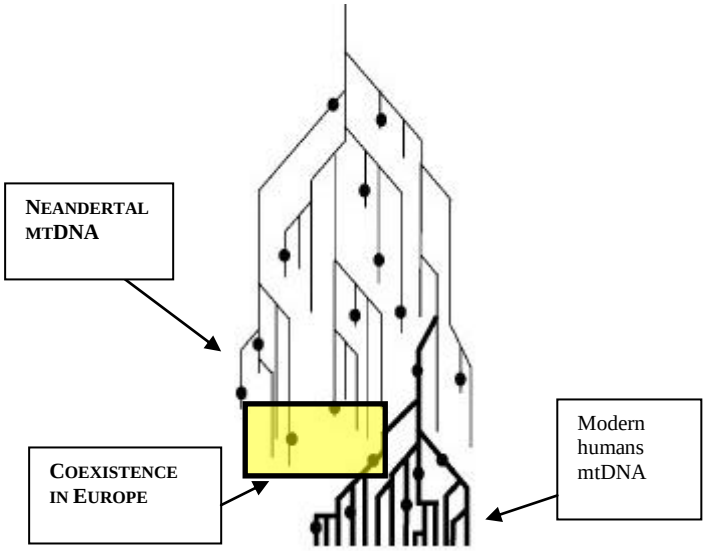


Fig. 5.4:3. Coexistence of Neandertals and Upper Paleolithic anatomically modern humans in Europe
Rys. 5.4:3. Koegzystencja Neandertalczyków z ludźmi anatomicznie współczesnymi górnego paleolitu

Based on study reported in Cyran and Kimmel (2010) (see also section 5.3), the Wright – Fisher model is equivalent to the branching process with the number of offspring having the Poisson distribution – then the distributions of the time to coalescence of a pair of sequences is identical to the O'Connell (1995) distribution (see Fig. 4).

The feasible value of α is 10 (see O'Connell 1995, Cyran and Kimmel 2010) and $\sigma^2 = 1.001 = E(\xi_0)$ for Poisson offspring number distribution with $\alpha = 10$ and $T = 10\ 000$. According to Theorem 3.6:3, equation (3.6:20), the probability of nonextinction of a linage

descending from a single Neanderthal mtDNA, $P(Z_t > 0 \mid Z_0 = 1)$ is given in the O'Connell (1995) model by

$$P(Z_t > 0 \mid Z_0 = 1) \sim \frac{2\alpha}{\sigma^2 T}\left(1 - \exp\left(-\alpha \frac{t}{T}\right)\right)^{-1}, \quad as \quad T \to \infty. \tag{5.4:1}$$

Consequently, the probability of extinction of such lineage is equal

$$P(Z_t = 0 \mid Z_0 = 1) = 1 - P(Z_t > 0 \mid Z_0 = 1). \tag{5.4:2}$$

Therefore, the probability of extinction of lineages started by x hypothetical mtDNAs present in the Upper Paleolithic *H. sapiens* gene pool is given by

$$P(Z_t = 0 \mid Z_0 = x) = 1 - P(Z_t > 0 \mid Z_0 = 1)^x. \tag{5.4:3}$$

The graph of the likelihood $P(Z_t = 0 \mid Z_0 = x)$ as a function of x is given in the Fig. 5.

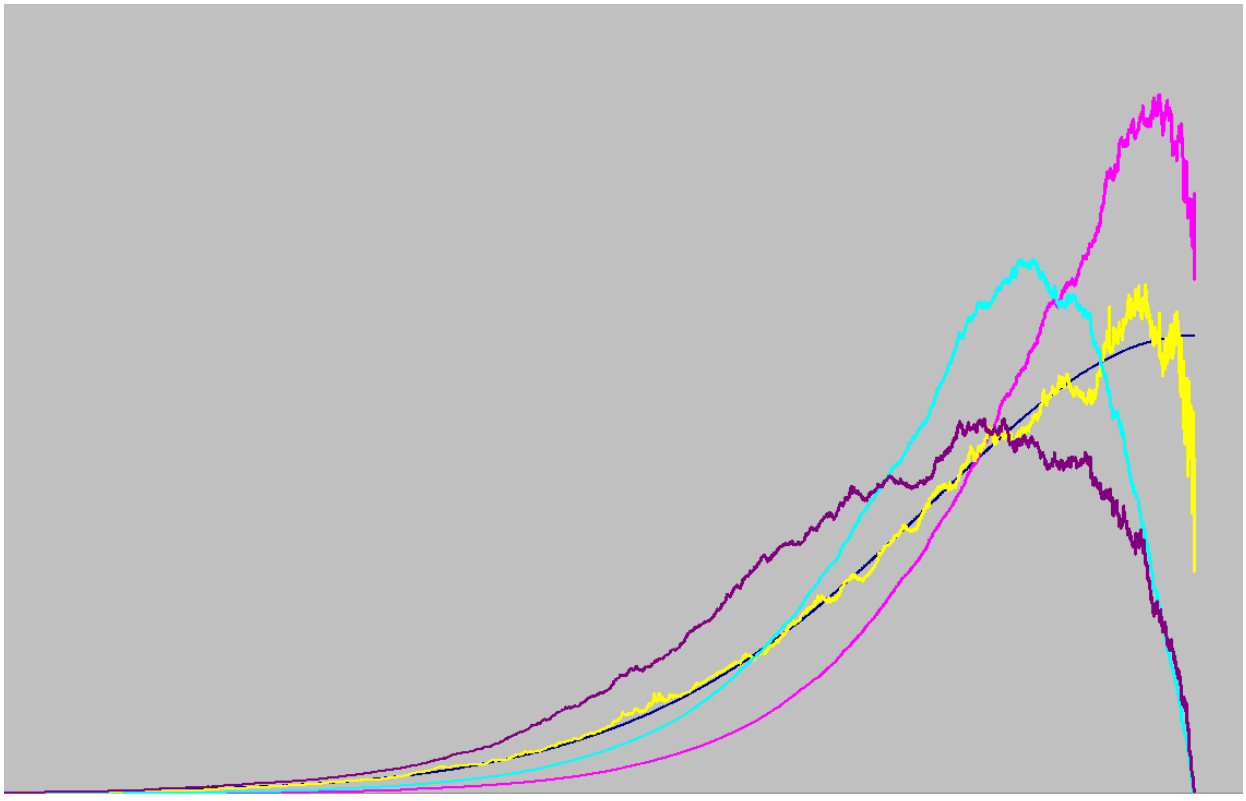


Fig. 5.4:4. Distributions of the time to coalescence of a pair of sequences
Rys. 5.4:4. Rozkłady czasu do koalescencji pary sekwencji

Solving (3) for x results in

$$x = \frac{\ln(P(Z_t = 0 \mid Z_0 = x))}{\ln(1 - P(Z_t > 0 \mid Z_0 = 1))}. \tag{5.4:4}$$

After plugging the data to (1), it follows that $P(Z_t > 0 \mid Z_0 = 1) = 2.57 \times 10^{-3}$. To compute the maximum admixture not contradicting the mtDNA record at 0.05 significance level, assume

the probability $P(Z_t = 0 \mid Z_0 = x)$ be 0.05. Therefore, from (4) it follows that the long-term effective population size x = 1,166 Neanderthal individuals.

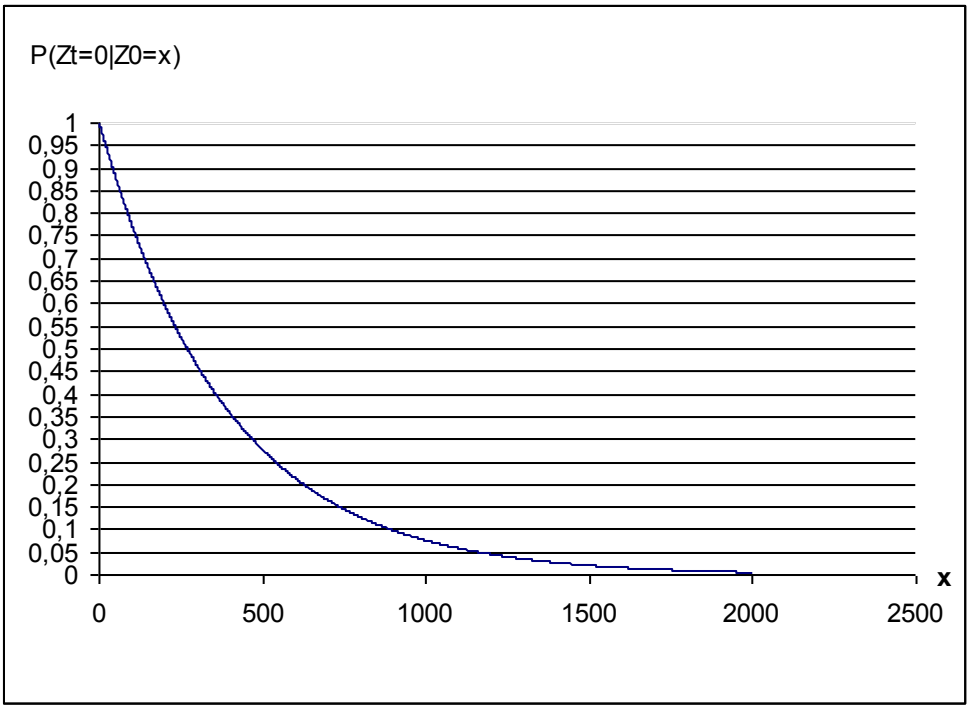


Fig. 5.4:5. The likelihood of the $P(Z_t = 0 \mid Z_0 = x)$ as a function of x
Rys. 5.4:5. Funkcja wiarygodności $P(Z_t = 0 \mid Z_0 = x)$ jako funkcja x

Using the Bayesian rule, the posterior probability is given by

$$P(Z_0 = x \mid Z_t = 0) = \frac{P(Z_0 = x)P(Z_t = 0 \mid Z_0 = x)}{P(Z_t = 0)}. \tag{5.4:5}$$

Assuming the uniform distribution of the prior probabilities $P(Z_0 = x)$, and an appropriate scaling factor $P(Z_t = 0)$ which is independent of x, it is possible to compute from (5) the distribution of $P(Z_0 = x \mid Z_t = 0)$. Having this distribution, the expected value $E(Z_0 \mid Z_t = 0)$ can be obtained as

$$E(Z_0 \mid Z_t = 0) = \sum_x xP(Z_0 = x \mid Z_t = 0). \tag{5.4:6}$$

It follows that $E(Z_0 \mid Z_t = 0)$ = 388 individuals, which is the most likely effective population size of the Neanderthal mtDNA sequences in the Upper Paleolithic *H. sapiens* mtDNA gene pool.

It is estimated that the census size of modern humans population around 30,000 years ago was at least 500,000. Therefore, the census population size of females active in reproduction at that time was at least 100,000 (assuming the same number of males and females in a population and provided that on average 1 out of 2.5 females in a population is reproductively active). Moreover, if the actual variance of the number of offspring σ^2 is 10 (that corresponds

to standard deviation about 3 – what is feasible), then, the effective short-time inbreeding population size of modern human females living 30,000 years ago, N_e is about 10,000.

To compute the expected value of the Neanderthal mtDNA admixture in a gene pool, let us divide $E(Z_0 \mid Z_t = 0) \approx 400$ by $N_e = 10{,}000$. This results in the expected admixture of about 4%. Similarly, to compute the maximum admixture non contradicting the mtDNA testimony at significance level 0.05, let us divide $x \approx 1{,}200$ by $N_e = 10{,}000$. This results in the maximum hypothetical admixture of about 12%. Both, above estimates are corroborating with the latest results obtained based on nuclear DNA sequenced from the Neanderthal fossils in the Neanderthal genome project (see Green et al. 2010).

5.5. Conclusions

It is a well known fact that the results of the search for natural selection operating at molecular level are affected by population history. Therefore the estimation of the probable long-term demographic history of a population, and in particular, the detection of the past population growth has become one of the main problems in statistical genetics. In the last decade, with the advances of new numerical methods and the more and more productive computers the forward-time simulations (described in section 5.2, see also Cyran and Myszor 2008a) started to play the role reserved earlier for coalescent methods.

On the other hand, artificial neural networks have been successfully used for years in many scientifically sound problems. Neural networks have ability for adaptation and generalization of knowledge, and can find hidden patterns in input data by inductive machine learning process (see section 2.2.1). Therefore they might be successfully used in solving problems that are often hard to describe by rule-based algorithms, such as those presented in section 2.4 and 5.2. The crucial is point is the availability of the training data representing properly the problem considered. Section 5.2 describes how such training sets for detecting past population growth were obtained by forward-time simulations.

Detection of the past population growth is one of the crucial issues in contemporary population genetics, particularly with regard to the human populations evolution, described in section 5.1. The importance of the problem is especially well understood in the context of neutral theory of evolution at molecular level proposed by Kimura (see section 4.1). This theory often serves as neutral hypothesis in the search for genes which underwent natural selection (section 4.3). The conclusions in such studies can be false if population expansion was present but not detected and therefore not introduced into the model (for example using author's MNH method).

In the studies dedicated for the detection of population growth the researchers often use various statistics computed for the same sample and then they try to analyze the results and draw conclusions (Reich and Goldstein 1998, Reich et al. 1999, King et al. 2000, Fisher et al. 2001). For example, the role of dominant mutations present in population which underwent the expansion period was analyzed by Cebrat and Pekalski (2004). The goal of the AI-based method presented in section 5.2 was to create a test which would be able to encompass knowledge gained from several known statistics based on microsatellites. Such novel statistical test γ, which emerged from application of artificial neural networks theory and was designed to detect past population growths based on genetic microsatellite data (see also Cyran and Myszor 2008b) is a confirmation of usability of the proposed AI-based methodology.

In experimental part of the research, there were created sets of samples, using forward-time simulation methods, described in section 5.2. These samples were picked at random from simulated populations that had undergone growths of different types and intensities. Then, different artificial neural networks were created, trained and the power of new tests based on these networks was experimentally verified. Finally, the comparison was performed of a power of the author's new ANN-based γ test with powers obtained by known methods based on microsatellites. Studies performed by Cyran and Myszor (2008b) showed that the proposed γ test provides better power in detection of population growth than the best currently available tests based on microsatellites, such as Kimmel's and King's imbalance indices (King et al. 2000).

The growing interest in studies concerning genealogies of branching processes is reflected among others by studies of Klebaner and Sagitov (2002) focused on geometric distribution of progeny, or work of Lambert (2003) dedicated for subcritical cases. Nevertheless, in the book the O'Connell model was considered as a standard because its independence of offspring distribution and the interest in supercritical processes dictated by an observation of long-term growth of human population size.

Contrary to the O'Connell model the Wright-Fisher model is not limited to any specific growth patterns and for historical reasons and simplicity, it is accepted in many methods of analysis of genetic diversity. Yet, except for some early classics, like for example Nagylaki (1990), relatively little effort has been expended in analysis of its relationship to other models and robustness to estimating errors caused by departures from model assumptions. Addressing this problem, there were compared in section 5.3 coalescence distributions under a range of Wright-Fisher models including those which arose from time continuous coalescent as well as distributions obtained from the O'Connell model.

Finally results of all these models were compared with actual distributions obtained from simulations of several thousand full genealogies using designed by the author computer software. Even if further detailed analysis of that fact is beyond the scope of the monograph, it is worth it to notice that implemented by the author simulation-based approach with full genealogy is capable also for computations of actual coalescence distributions of a pair of alleles or estimates of the time to MRCA of the whole populations in virtually all types of population evolution. On one hand it is not limited to multinomial sampling, like Wright-Fisher models, on the second it is not limited to homologous in time branching processes like O'Connell model. As a real, biologically sound application of these results, in section 5.3 there were reported estimations of the time to mitochondrial MRCA of modern humans showing how insensitive is the estimate to considered models' assumptions.

The Wright-Fisher model of genetic drift assumes a panmictic population. However, it seems equally or even more likely that modern humans colonized new territories in small isolated groups, which were frequently becoming extinct. This latter pattern seems more similar to a branching process. Do these two different models of population dynamics lead to radically different estimates of the age of the mtEve, or any other common ancestor? In this chapter, there was presented an attempt how one might answer to this question using intensive computer simulations and comparing it with known genetic models.

Until the last decade, the estimation of the divergence rate could rely only on the human-chimpanzee divergence data. Methods used were based on phylogenetic trees constructed either by maximum likelihood or parsimony and rooted using chimpanzees as an outgroup. However, due to relatively long time to this divergence, all estimates of this time were very inaccurate, this latter ranging from 4 to 9 million years. Consequently, estimated divergence rate and the time to MRCA of modern humans could not be accurate, with expectation ranging from 200,000 years ago (Wilson and Cann 1992, Vigilant *et al.*1991) to 300,000 years ago (Hasegava and Horai 1991). Additionally, in population genetics, many possible patterns of the human population growth were assumed. The simplified exponential models were often used, but also the logistic growth of human population proved to be not inconsistent with the mtDNA variation data (Polański *et al*. 1998).

Mentioned above predictions were in agreement with the out-of-Africa scenario and in contradiction to the multiregional theory of origin of modern humans, supported by some paleontologists (Thorne and Wolpoff 1992). These researchers claim that the time to MRCA should be placed about million years ago or even earlier. And, what should be emphasized here, genetic data did not necessarily contradict the multiregional theory, as it was shown by O'Connell (1995). He inferred, using the branching process model, that the genetic diversity of modern humans was consistent with estimates of the mitochondrial Eve epoch between 700 thousand and 1.5 million years ago.

These estimates depended on inaccurate inference of the human-chimpanzee divergence time and on the methods of inference used. To validate his conclusions, O'Connell (1995) also indicated the weak points of the outgroup methods when the outgroup was not close enough in genetic distance to the sample. If application of different methods to the same genetic data had given results differing by almost one order of magnitude, the multiregional hypothesis could not have been rejected solely because it was in contradiction with the majority of genetic methods, while there were still methods supporting it.

Situation has changed after 1997 (Krings *et al*. 1997), when for the first time the mtDNA from *H. neanderthalensis* dated to live until about 40,000 years ago (Schmitz *et al*. 2002) was sequenced. However, only less than 400 base pairs were sequenced, hence any estimates based on this data were not very reliable. The next successful sequencings of Neanderthal mtDNA in 1999 (Krings *et al*. 1999) and in 2000 (Ovchinnikov *et al*. 2000, Krings *et al*. 2000) confirmed the accuracy of the first experiment and qualitatively changed estimation of the time to the most recent common female ancestor of modern humans, which now no more solely relies on the dating of human-chimpanzee divergence event. Since it seems more probable from genetic data (Krings *et al*. 1999) that *H. neanderthalensis* did not contribute mtDNA to modern humans, the time of the mtEve should be placed after the *H.* sapiens – *H. neanderthalensis* divergence. Even if later studies (Serre *et al*. 2004, Cyran and Kimmel 2005) indicated that interbreeding between the two human forms could not be excluded it remains true that the root of mtDNA of living humans should be after that of humans and Neanderthals.

The hypothetical admixture of at most 25 % (Serre *et al*. 2004) or 15 % (Cyran and Kimmel 2005) disappeared in the process of genetic drift. Therefore, even if some researchers using early results of Neanderthal Genome Project suggest possible interbreeding between Neanderthals and archaic Europeans, yielding at least 5 % admixture of nuclear DNA (Plagnol and Wall 2006, Pennisi 2006), the methodology used in the section 5.3 based on treating Neanderthal as an mtDNA outgroup is well justified. In the context of discussion about interbreeding, it is also worth to mention that results of Neanderthal Genome Project interpreted by other scientists lead them to conclusion similar to those formulated based on mtDNA, i.e., they suggest no Neanderthal admixture in modern humans gene pool (Noonan *et al*. 2006, Pennisi 2007).

Section 5.3 of this chapter compared distributions of the time to coalescence of a pair of alleles obtained by conceptually different methods. In particular it is shown there that branching process evolving for as little as 100 generations yields the O'Connell asymptotic coalescence distribution expectations, which differ from the actual expectation computed in the full genealogy model by less than 2 %. Moreover, this result holds for any offspring distributions and due to the asymptotic character of the O'Connell results it remains true also

for realizations of branching processes with arbitrary large number of generations. Having this important result, it is possible to obtain the expectations of the ratio of the coalescence time of two individuals and that of all individuals in the population also for numbers of generation of the order of 10^4, even if it is infeasible to apply the full genealogy model in this case.

At the end of section 5.3, proposed approach was applied to estimate the age of the most recent female human common ancestor, based on the genetic material from hypervariable control regions I and II of the mtDNA belonging to contemporary humans and Neanderthal fossils. For all stochastic trajectories analyzed, the resulting time falls into the 95 % confidence interval of the estimate based on the phylogenetic trees (Krings *et al*. 1999). Yet, the results presented here, based on expectation obtained in the full genealogy model equal to 126×10^3 years, indicate a shift of around 30×10^3 years towards the present, as compared to Krings *et. al* (1999) phylogenetic tree-based estimate, equal to 163×10^3 years. However, since this shift is relatively small (23 %) and it is mainly the result of the assumed difference in time of the split of Neanderthals and modern humans, rather than different methods applied, one can conclude that the stochastic models based on branching processes provide similar estimates to those, based on phylogenetic analysis, therefore supporting each other.

Therefore, the results described in this book indicate that the estimates of the time to coalescence in the Wright-Fisher and in the coalescent models are quite robust. They deviate by less than 8 % (see Table 5.3:8) from the standard O'Connell model predictions, whereas the asymptotic O'Connell expectation differs from the actual expectation, computed in the full genealogy model, by only 1.6 %. Such small differences are in a clear opposition with large range of confidence intervals obtained not only in pairwise difference based methods considered in section 5.3, but also in the phylogenetic studies. The greatest level of uncertainty about the expectations are caused by such scaling factors like between-species divergence rate and not by deviation from particular assumption of the method used. This validates both the Wright-Fisher and the coalescent models also for population histories not following assumed within these models scenarios. In particular, it validates results about inferring population trajectory from the genetic diversity data, as reviewed in Wooding and Rogers (2002), results which implicitly relied on the Wright-Fisher model assumption, but which remain valid for much larger spectrum of possible demographies.

As long as some known facts are difficult to understand in the light of given hypothesis, the alternative hypothesis cannot be neglected. That is why, instead of trying to disprove multiregional or out-of-Africa model, both having troubles in explaining some known facts, we draw conclusions which can make any of these models more reliable. However, the consequences of our inferring are not equal for these two (still alive) competitive hypothesis.

For out-of-Africa scenario our results quantitatively show to what extent it can rely on mitochondrial DNA (mtDNA) inferences. It is important in drawing conclusions, when morphological fossil record possibly contradicting the pure version of out-of-Africa hypothesis would have been discovered. In other words, even if to-date mtDNA-based results do not contradict the radical form of the recent out-of-Africa origin, relying solely on them, cannot be treated as a prove of the model of total replacement. Such total replacement of archaic *Homo* populations by descendents of mitochondrial Eve without any admixture from archaic autochthons gene pool, is unlikely in the light of paleoanthropological fossils, as it is often emphasized by multiregionalists.

On the other hand, indicating correctly the insufficiency of mtDNA-based inferences, is not equivalent to ignoring them and treating mtDNA and molecular clock based methods as a source of completely unreliable information. Since mtDNA, recently sequenced from Neanderthal and ancient *Homo sapiens* fossils, can be the base for estimation of the upper limit of plausible Neanderthal mtDNA contribution to descendants of Eve, the multiregional model should postulate assumptions not exceeding dramatically these limits. Such limits of admixture have been lately computed to be about 25 percent, and have been presented in (Serre et al. 2004), but Cyran and Kimmel (2005) further reduce the extent of plausible Neanderthal mtDNA contribution applying another strategy to 15% (see also section 5.4, for the most recent results, which reduce this estimate to 12%).

The method based on a branching process allowed to calculate the limit of Neanderthal mtDNA admixture which (with probability 95 percent) would have been preserved until present if it had been really added into human gene pool some 30,000 years ago. It is the time when the Neanderthals have probably disappeared, but the question arises whether and to what extent they contributed to modern humans gene pool before extinction.

Thorne and Wolpoff (2003) suggest that (a) this contribution should be up to 50 percent in an early population of modern humans in Europe, and this is only the genetic drift that cleared any trace of this fact in contemporary humans. These authors also claim that (b) mitochondrial inferring concerning Neanderthals is unreliable due to contamination of mtDNA from Neanderthal fossils by contemporary sequences. At the opposite site, some supporters of out-of-Africa model, treat the mtDNA testimony as (c) the evidence of no Neanderthal mtDNA contribution at all. Below there is short discussion of these issues comparing with the results presented in section 5.4.

a) The author's results show that Thorne and Wolpoff (2003) present extreme but not necessarily false opinion concerning the amount of Neanderthal admixture. Our conservative assumptions led us to a maximum level of admixture being about 15 percent with 95 percent confidence. However this estimate was calculated as the proportion of

Neanderthal mtDNA in the whole human population. Thorn and Wolpoff (2003) are talking about proportion in Europeans. The change from 15 percent in total population to 50 percent in Europeans, as postulated above by these multiregionalists, demands that Europeans constituted at most 30 percent of the whole population. This seems feasible and therefore admixtures of similar magnitude can be accepted in the light of the to-date knowledge. However we must stress that we have used in our calculations the minimum estimate of the human population size 30,000 years ago. The maximum estimate is more than 10 times larger (Jobling et al. 2004), and if this second estimate proves more likely, then estimated here maximum Neanderthal admixture will correspondingly decrease ten times, disproving claims about 50 percent Neanderthal admixture in early Europeans.

b) The possible contamination by modern sequences of ancient Neanderthal DNA would not probably yield sequences noticeable different from contemporary modern humans. Yet it did, so if they were really contaminated, it means that in reality the genetic distance between *H. sapiens* and *H. neanderthalensis* would have been even greater than that estimated during Neanderthal sequencing studies. Therefore it is hard to understand why this argument is raised by multiregionalism supporters.

c) The percentages of Neanderthal mtDNA admixtures which cannot be excluded based on sequencing studies, estimated by Serre et al. (2004) to be about 25 percent, should not be treated as the evidence of no contribution at all. Yet, some important papers (mainly the older ones, like Krings et al. 1997, 1999, Ovchinnikov 2000) when announcing the fact that contemporary mtDNA gene pool does not contain mitochondrial genes inherited from Neanderthals, seemed to neglect the effect of genetic drift, what (together with the overstating the conclusions drawn from the lack of regional affinity of Neanderthals with contemporary Europeans) was criticized by Relethford (2001).

The results of Cyran and Kimmel (2005), although further reduce the plausible maximum amount of admixture to 15 percent, still cannot be used as a proof of no admixture, but on the other hand do not contradict such radical form of replacement. The most recent results of the author (see section 5.4) which suggest that the 4% admixture of Neanderthal mtDNA was present in the mtDNA gene pool of anatomically modern humans some 30,000 years ago, do not change the above conclusion.

6. EARLY LIFE

6.1. Foundations

First traces of life on the Earth are 3.5 billion years old (Orgel 1998), but it is commonly assumed that life began 0.5 billion years earlier. At the beginning, life on our planet was completely different from the present one. There is a lot of unanswered questions connected with this period. We do not know where did life start (Edwards 1998, Trevors 1999): on the Earth's surface, in deep-sea vents (Orgel 1998) or maybe, as panspermia advocates suggest (Hoyle and Wickramasinghe 1999), it came on Earth in meteorites, which were common guests on our young planet, in this period. We do not know also what was the first: metabolism or replication (Pross 2004), or maybe metabolism and replications emerged in the same moment.

We wonder whether the RNA-world (a world in which RNA molecules were the only available form of life, which led to DNA creation – according to this theory, RNA strands acted both as information carriers and catalysers of chemical reactions), was the first one or maybe was it preceded by other forms of life, such as based on peptide nucleic acid (PNA) PNA-world, based on threose nucleotide analogs (TNA) TNA-world, based on pyranosyl analog of ribose (p-RNA) p-RNA-world, based on alanyl nucleic acids (ANA) ANA-world, or based on glycol nucleic acids (GNA) GNA-world (Joyce and Orgel 2006).

As simple consequence of the observation that life exists is the fact that there must have been some beginning of it. Despite many evidences that RNA world existed and predated current life based on DNA there is still a lot of unanswered questions and troubles to be solved. A vast number of experiments relating to life's beginning are currently carried out by computers. Constant raise in computational power of those devices, let us create and analyze more and more sophisticated models and elaborate conclusions derived from the older ones.

Computer Monte Carlo simulations of RNA world include different stages of the origin of life. Some models, such as those considered in section 6.2 and 6.4, rely on real chemical processes operating on RNA strands. In particular, model described in section 6.2 takes into

account the phosphodiester bond break process, which leads to the hydrolysis of the RNA strands, and in section 6.4, the focus is on the non-enzymatic template-based RNA recombination processes. According to many researchers these latter processes might lead to an elongation of RNA chains and creation of novel sequences in the solution. These new RNA chains could have catalytic activities and serve as RNA replicase. The results of the study described in section 6.4 let us conclude that RNA non-enzymatic template-directed recombination processes are important phenomena in the RNA world and might lead to significant RNA chains elongation, required for the emergence of the primordial RNA replicase, which however, was the subject to the length restrictions, as given in section 6.2. The problem of how many other genes (additionally to the replicase, which seems to be thm most crucial gene) could have existed in a primitive compartment-based protocells is a subject of the study considered in section 6.3.

Before continuing the description of the scientific views on how biological life has emerged, let us focus on some fundamental issues involved in self-replication process, considered from a perspective of technical sciences. It is a fact, that in some aspects, living organisms resemble self-replicating automata. Therefore, before giving the description of the bio-chemical theories of the origin of life, let us consider the structure of the universal automaton proposed by Turing (1936, 1938), and the concept of self-replicating automaton, described by von Neumann (1951).

Definition 6.1:1 (Turing machine, after Turing 1936, 1938)

The Turing machine is an automaton which can be in one of the states S_i, where $i = 1, 2, \ldots, n$ and n is an arbitrarily large but finite integer number. The communication with external world is performed by the tape composed of symbols corresponding to logic zero and one. The automaton can read or write only one symbol e at a place under direct inspection and is able to move the tape in both directions by one position. The operation of such machine is described by the sequence $S_i(t)$ of states at discrete time events t. The transition to given state $S_j(t+1)$ from a state $S_i(t)$ is accompanied by the shift of the tape by p positions, where $p = -1, 0, 1$ and inscription on the tape a symbol b, where $b = 0, 1$. Therefore, the complete definition of functioning of such an automaton is specified by functions $S_j = f(S_i, e)$, $p = g(S_i, e)$ and $b = h(S_i, e)$.

▬

Since (possibly infinite) sequence of bits can be treated as binary expansion of the real number, the operation of presented automaton can be considered as a process (possibly infinitely long) of computation of that number. Hence, Turing has solved the problem of the structure of universal automaton U, i.e., such automaton which (given sufficiently long time) is be able to produce any sequence which can be produced by any other automaton X. In other

words universal automaton is able to produce arbitrarily long part of (possible infinite) sequence of bits representing any real number. Turing proved that if the (finite) starting part of the tape is considered as instructions for the universal automaton *U* and assuming that the law of forming some desired sequence is known (i.e., assuming that we know the definition of some other automaton *X* capable to produce the desired sequence which should be implemented also by *U*) it is possible to express such instructions in a form of finite sequence of zeros and ones forming program for universal automaton *U*. The processor of such instructions in universal automaton must only be able to implement functions *f, g, h* defined above.

Turing observed that the complete general description of any conceivable automaton can be expressed in a finite number of words consisting also empty passages corresponding to functions *f, g, h*. As long as they are empty, the schema represents general definition of any automata – it becomes specific after filling them with desired functions. Now let us imagine an automaton *U* capable for the interpretation of such schema – it defines Turing's universal automaton.

Definition 6.1:2 (Turing universal automaton *U*, after Turing 1936)

The automaton *U* capable to imitate the operation of some other automaton *X*, i.e., automaton which, when fed with the definitions of the functions *f, g, h*, for automaton *X*, can operate like the automaton *X* described by these functions and to imitate the operation of object described, is called the Turing universal machine. Such machine can duplicate any action of any conceivable automaton *X* when it is fed with description of that other machine. The description of *X* is software for *U*, i.e., *U* is the hardware interpreter of *X*.

▬

Note that, any automaton *X*, encoded as software program, can operate only if it finds some hardware automaton capable for interpretation of it. This clear logical precedence of the hardware over the software is relevant also for biological information processing systems, especially in the context of the origin of life. It is also visible in the case of self-reproducing artificial automata described by von Neumann (1948) and considered below, after addressing one of the most fundamental problems, which arises when asking and trying to answer the question: what is life.

This is a problem of components and processes which constitute life. Among scientists, there can be found two kinds of view which incorporate many different in details answers. The first is the opinion that life is almost the same as replication. This is probably most widely accepted outlook, especially sound among researchers studying problem of the origin of life. The proponents of this view treat metabolism as biologically important factor but not

as a *conditio sine qua non* of life. The second, idea assumes that for life to be present, not only replication, but also metabolism is essential.

Some of the replicative life advocates claim there is an experimental verification of their view at least as old as the discovery of viruses. However, viruses, i.e. purely replicative creatures can be considered as the extreme case of parasites for metabolic cellular life. The fact that viruses do not metabolize is not a convincing argument against necessity of metabolism in life. At most it shows that not all living organisms – although some researchers (Cajavec 2002) do not consider viruses as living creatures – must metabolize. At first sight it may look like the proof of the life without metabolism, but in fact it is not. Even if viruses are regarded as living things, when trying to prove the only-replication view, there is a need to take into account that without metabolizing cells viruses would not replicate. Hence for life as we know it and treated as a whole system, the metabolism seems to be crucial and irreducible.

This discussion if a foreground for the studying the logical connections between replication and metabolism. As it will be shown, it is possible to measure the information in the metabolic system. So finally, this discussion serves as the excuse for including to the book written from the information processing perspective the issues of metabolism. Contrary to that, the importance of information theory for the replication of genomes, i.e. structures composed of extremely long words coded with the use of four symbols alphabet does not require further explanation.

After all, these introductory steps let us consider the old but fruitful work of von Neumann. In his famous talk in 1948 (reported in Von Neumann 1951) he introduced the concept of self-reproducing universal automata. Although he obviously knew that "*Natural organisms are, as a rule, much more complicated and subtle, and therefore much less well understood in detail, than artificial automata*", nevertheless, he also stated that "*some regularities which we can observe in the organization of the former may be quite instructive in our thinking and planning the latter, and conversely, a good deal of our experiences and difficulties with our artificial automata can be to some extent projected on our interpretation of natural organisms*" (von Neumann 1951).

The concept of automata which can build copies of themselves is the extension of the described above Turing universal machine *U* (Turing, 1936). The obvious limitation of all Turing machines is that they are purely computational automata – they can produce nothing but sequences of bits. In fact this is serious limitation in the context of self-reproduction, but nevertheless, it is worth to notice that ability to produce arbitrary sequence of bits by *U* machines puts them not too far from the goal. Their products (i.e. sequence of bits) can be considered as software describing operation of identical or similar automata to *U* itself.

Therefore, one universal automaton can produce a lot of such software programs defining other universal automata.

Yet, it cannot reproduce itself, because it is incapable of producing hardware of which it consists. Hence, described here system cannot be a logical basis for the life. Despite the increasing number of software-based (resembling viruses) individuals (and perhaps even species), the system will surely collapse when the individual resembling automaton *U* dies. The existence of many copies of the programs describing operation of *U* will not help as long as processor interpreting these programs is lacking. The clear problem is that such individual cannot reproduce its hardware.

What should be added to *U* to overcome this limitation? Obviously the output of self-reproducing automaton should not be a program describing any automaton but the automaton itself. That is hardware and software. Or hardware only in the case of purely hardware and thus non-universal, yet functioning automata. But definitely software only is not a sufficient output. Von Neumann showed that results similar to those obtained by Turing can be extended to automata producing other automata. The following theorem regarding self-reproduction and its proof is based on the lecture given by von Neumann (1951).

Theorem 6.1:1 (Self-reproducing automaton, after von Neumann 1951)

It is possible to build self-reproducing automaton $H_S(I_S)$ (composed of hardware H_S and software I_S describing H_S) such that, in the appropriate environment supplying it with necessary components, it will produce copies of $H_S(I_S)$.

Proof

It is clear that it is possible to describe general automaton whose output is any another automaton. Of course such description must have empty spaces which should be filled with the description of particular structures and functions of the specific automaton to be produced. Let automaton *P* be such machine which when fed with the description of any other automaton *X* and an sufficient supply of elementary parts will produce that object. The description of *X* is called the instruction *I* and it is given as a combination of structural elements of *X*. It is worth to notice that in general the complexity of *X* can be greater or lesser than that of the constructor P because the complexity of *X* is determined only by the instruction I. In other words *X* can be very complicated structure or it can be arbitrarily simple object. However, complexity of *P* cannot be arbitrarily small. It must be sufficiently large for allowing *P* to produce *X* according to instructions *I* with adequate quality.

Consider also automaton *R* capable for reproducing any instruction *I* and automaton *C* representing control mechanism for combination of *P* and *R*. At first, *C* orders *P* to produce new automaton according to *I* and orders *R* to produce copy of *I*. Then it transfers the copy of *I* to the automaton produced by *P* and finally it releases the construction from the ensemble

$P + R + C$. Let the ensemble $P + R + C$ together with instruction I constitute automaton H_S (I) composed of hardware H_S and software I. Finally, let the instruction I have the specific form I_S which describes H_S. Then the automaton H_S (I_S) composed of hardware H_S and software I_S is clearly self-reproductive, what ends the proof.

■

Implication: Observe that vicious circle has been avoided because I_S describes only hardware H_S (to which it is only added without modification) and it therefore does not describe complete self-reproductive automaton H_S (I_S). Therefore, this result determines the chronological, as well as logical, precedence of hardware H_S over software I_S in self-reproducing automata (whether artificial or biological). The H_S has to be formed before the construction (i.e. copying) of I_S is invoked and only then "*the process is legitimate and proper according to the rules of logic*" (von Neumann, 1951).

Such automata together with the ample reservoir of the components constituting its hardware can form a self-sustaining system of self-reproducing machines. Moreover, such artificial systems not only can self-reproduce but also can evolve. They can be used as a logically consistent analogues of living organisms. For example it is evident that the function of automaton R within self-reproductive machine is equivalent to the replication of the genetic material within the living cell. Therefore it can be easily imagined what will happen if there is non-zero error rate in the functioning of automaton R. The phenomenon of mutation, occurring as the result of it, in vast majorities of cases is deleterious. In some cases however it can be responsible for new traits which could turn out to be advantageous especially in changing environment.

Such mutations, responsible for the evolution of life, can be even better modeled if a small variation to a foregoing construction will be added. Let us imagine the automaton H_S ($I_S + E$) composed, like before, of hardware H_S, however with software $I_S + E$ describing not only its self-reproductive hardware H_S but also some additional structures E analogues of enzymes not involved in a reproduction. If in such automaton the mutation occurs in the E component of the instruction $I_S + E$ then it will not be lethal for the reproduction cycle. Instead it will produce new self-reproductive automaton H_S ($I_S + E_1$) subject to natural selection considered by Darwin as the leading force in the evolution of natural life.

Then, what is life? How did it originate? Is replication required for reproducing, heredity, and thus natural selection? Can replication and metabolism be (at least logically) separated? Let us consider logical relationships between replication and metabolism. Such questions as whether these two processes can be separated (leading to imagination of replicative life without metabolism or metabolic life with no replication) must be asked to avoid a trap of taking for granted any particular view.

Dependently on the answer, the origin of life can be considered as subsequent origin of these two phenomena or simultaneous occurrence of both. The latter is referred to as hypothesis of a single-origin, the former (at least from a formal point of view) always is a case of double-origin hypothesis. However, for reasons explained below, only one particular form of double-origin hypothesis is really referred as such. It is hypothesis in which metabolism appears before replication – the opposite is still referred to as single-origin, despite formal stipulations.

In theories assuming the sequence: *first metabolism – then replication*, the origin of self-replication is explained similarly as in *first replication –then metabolism* theories, however, with one crucial difference. This difference of a great importance concerns the environment of self-replicating macromolecules, precursors of modern genes. This issue, namely the problem of occurrence of replication in already biotic conditions versus origin of replication in pre-biotic environment will be discussed in more detail in section 6.2. This section will try to support the reader with arguments for treating this difference as really relevant and meaningful, based on results of experiments performed by the author (see Cyran 2007b, 2008a) using the Demetrius/Kimmel branching processes model (Demetrius et al. 1985, Kimmel and Axelrod 2002).

These experiments, suggesting substantial limitation of the complexity threshold in the early life, although by no means decisive, gave favor to the origin of replication in biotic conditions by indication how difficult, based on contemporary biochemical experiments of RNA molecules evolution in a test tube, it is to imagine self-replication with required quality in pre-biotic environment. Perhaps it is worth to notice that these conclusions corroborate with results obtained with the use of others methods (based on balance between information loss and selection) by equally clear indication of difficulties with the origin of self-replication in abiotic environment.

Nevertheless, the view that life originated twice in a sequence: first metabolism and then replication, is not popular. More fashionable view, perhaps due to its elegant simplicity, states that life started with replication. The origin of metabolic apparatus within such theories is not considered as the second origin of life, rather it is treated as the step (milestone) in the evolution. It is very neat picture which is nowadays represented by many theories of RNA world. This term, introduced in 1986 by Gilbert (1986) and further by Joyce (1989), refers to hypothetical early stage of the evolution when both genetic (which occurred at first) and structural/metabolic (which occurred later) functions were realized by molecules of RNA.

However, the estimated conditions required for avoiding the error catastrophe, which could be caused by the loss of information in the RNA world, the results of computer simulations performed by Niesert et al. (1981), which proved instability of hypercycles proposed by Eigen et al. (1981), as well as outcomes of computer simulations of metabolic

self-sustaining cycles done by Sagre and Lancet (1999), suggest, that perhaps this neat picture should be changed to a sort of garbage-bag world. Garbage-bag world is a term coined by Dyson (1999) and represents life after the first beginning – beginning of metabolism and before the second origin – origin of replication in already biotic conditions. As it was already mentioned, the author's numerical experiments with criticality and extinction of branching processes seem to confirm the Dyson's view.

In all examples of known to us cellular life the relation between replication and metabolism is defined by the inherent circular interconnections. The replication of the genome is possible only with the use of protein enzymes catalyzing metabolic activity of the cell and the production of protein enzymes is directed by the information encoded in a genome. In such circular dependencies it is impossible to separate one phenomenon from the other, and hence it is hard to say directly which function originated first. However, despite some difficulties, it is possible to imagine other forms of life which are logically coherent, but they can also provide a clue about the sequence of events leading to modern life.

The rationale for such thought experiment is guided by the conjecture that two unlikely events (such as the birth of replication and metabolism in abiotic conditions) are more probable to occur sequentially than at the same moment. This is even more true if we take into consideration that the second unlikely event would occur then in already biotic environment organized by the first. Nevertheless, assuming that both, proteins performing metabolic activity required for replication of the genome and the genome composed of nucleotide strands required for coding and production of such proteins, occurred at the same time in appropriate neighborhood area separated from the external world by a sort of lipid barrier, then the single-origin of life is a simple consequence of this assumption. Yet, minority of scientists (if any) believe it to be possible. Rather, they suggest separation of the events in time, with majority advocating the birth of replication before the start of metabolism (Gilbert 1986, Szathmary and Demeter 1987, Joyce 1989, Smith and Szathmary 1999, Mc Ginness and Joyce 2003) and minority believing the contrary (Dyson 1999).

As it was already said, the majority views will be referred here, after Dyson, single-origin theories. Despite some formal stipulations, such name is acceptable, because the "second" start in these theories, that of metabolism, is described within them in terms of evolution of self-replicating system of life. As being only the result of evolution, it is considered not equally important as compared with the "first" origin, i.e. origin of replication. Consequently, the origin of self-replication is considered not only as the first origin, but as the only origin, implying propriety of the single-origin terminology. Contrary to above, the minority view is that the first sort of life started with the birth of metabolic (autocatalytic) activity within a proto-cell and that the second origin (origin of different type of life – replicative life)

appeared when self-replicating nucleotide strands became the parasites in their host protocells.

In such double-origin hypothesis there are two distinct forms of life: autocatalytic (but not replicating) life of host protocells and purely parasitic life of self-replicating (but not metabolizing) nucleotide macromolecules. The evolution of both kinds of life during subsequent millions of years led by the symbiosis to the complete interdependence, visible in contemporary life where it is impossible to separate physically the two processes.

After this discussion, with a goal not to treat it for granted that the life originated with replication, two next issues should be clarified. The first is the difference between replication and reproduction, the second is the meaning of the word metabolism. Very often reproduction and replication are used as synonyms, since always in observable life reproduction of the cell is performed with the replication of the DNA (or RNA) molecules in genomes. Yet there is a fundamental difference between the two: reproduction is an action of the cell dividing into two daughter cells with similar properties.

Even if today this process is always accompanied by the replication of a genome, the latter is not *conditio sine qua non* of the first. Equally well we can imagine reproduction performed in more statistical fashion – it would yield of course system with lower level of inheritance but it would not suppress inheritance all together (Dyson 1999). And it is inheritance produced by reproduction and not the replication of macromolecules what is important for the Darwinian selection to operate in the evolution. In fact, Darwin had no idea about replication of nucleotide polymers in chromosomes when he proposed his theory of evolution. Therefore, replication is not an assumption, neither is it the consequence of his theory. Rather, it can be treated as a very efficient (but by no means the logically sole) way of directing the reproduction leading to inheritance.

In all single-origin theories replication was the basis for reproduction from the very beginning. In the double-origin hypothesis replication of a genome evolved from parasitic replicative form of life long after its invasion on statistically reproducing proto-cells performing metabolic activities. Molecular biologists could be astonished by such conjecture as long as they mean the term metabolism as genetically driven activity of a cell. Yet such meaning, confirmed in all examples of contemporary life, is not the only one. Metabolism also means self-sustained autocatalytic activity of a cell capable for extraction of negentropy from the environment. Such meaning was prevailing in the times when the nature of the replication was unknown (Schro44) and it is still present especially in German language (Dyson 1999).

After clarifying the terminology, let us now consider in the three subsequent sections the three models of the early life: the branching process model used for estimation of the complexity threshold in the early RNA-world (section 6.2), random segregation compartment

model designed to estimate by forward-time simulations the maximum number of primordial genes before the organization of genetic material in chromosomes (section 6.3), and the forward-time simulation-based model describing the beginning of the RNA world from the organic compounds (section 6.4).

6.2. Complexity threshold

The amount of information in hypothetical RNA-protospecies can be considered in several stages of the RNA-World. Here, the interest is focused only on the early (but not the first) stage of this hypothetical world, i.e. the phase, which directly proceeds the first phase of short oligonucleotides of the length not exceeding 30-50 units (formation of such oligos from the nucleic acid components is considered in section 6.4). Note, that there is a radical difference between these two phases, which is manifested in possible ways of reconstructing the RNA polynucleotides after degradation. This issue will be explained in detail below, however, it should be stated now, that in both phases the protospecies are considered to be as simple as possible, i.e. they are single strands of RNA macromolecules.

Because of this very simple form of the protospecies in the considered stage of the RNA-world, the amount of information preserved in such organisms can be directly correlated with the length of the RNA strand. The four letter alphabet of adenine (A), cytosine (C), guanine (G) and uracil (U) is used to store the genetic information and the classical information theory can be used to compute the amount of information carried by this molecule.

However, with the length of the molecule the notion of the complexity threshold comes on the scene. This latter defines a maximum length of self-replicating RNA strands which could avoid error catastrophe. In this section there are summarized the main theoretical results of Demetrius/Kimmel model (Demetrius et al. 1985, Kimmel and Axelrod 2002) which can be used for computation of complexity threshold for different mutation rates and probabilities of RNA hydrolysis.

The novelty, which has been added by the author to this approach, is the introduction to the model the parameter associated with probability of the phosphodiester bond break (see Cyran 2009b), the parameter which can be experimentally measured in a test tube. Therefore presented here results can be easily refined after a series of biochemical experiments yielding the estimates of the feasible values of this probability under the geological conditions presumable present on the young Earth.

It is a well known fact that reproduction involving replication of genetic material produces almost identical copies of parental cells. The word *almost* is of great concern in the whole history of life since it reflects possibilities of rare changes caused by mutations on one

hand, and relative constancy of the genotype of the given species on the other. The exact replication of nucleotide strands could not have led to the whole variety of the living creatures. On the other hand, too large mutation rate would have led to error catastrophe and the process of evolutionary organization of life could not have proceeded.

Genetic experiments indicated that the rate of mutation in contemporary organisms is influenced by a lot of factors, such as the DNA repairs performed by protein enzymes called DNA helicases coded by such genes as RECQL, BLM, WRN (see section 4.3.1 for description of location and functions of these genes). They are involved in surprisingly many phases of DNA metabolism, including transcription, recombination, accurate chromosomal segregation, and various mechanisms of DNA-repair, such as mismatch repair, nucleotide excision repair, and direct repair. These mechanisms have evolved because genomes are often subject to damage caused by chemical and physical agents present in the environment, or by (Cyran et al. 2004) endogenously generated alkylating agents, free radicals, and replication errors.

Therefore, the effectiveness of the genome repair is one of the crucial factors determining the fitness of species. However, species capable of DNA repair must have long genomes used for coding many complex enzymes, including mentioned helicases. Hence, to assure more accurate replication (i.e. the smaller mutation rate), the longer nucleotide chain is required. This is of course reflected in the growing amount of information needed for the coding so many functions.

Yet, for longer chains there is smaller probability that they are replicated without error for given mutation rate due to almost independent replications of separate nucleotides in a chain. The conclusion of this discussion is the existence of maximum length of a poly-nucleotide strand that will not (almost surely, i.e. with probability one) become extinct. This length is called the complexity threshold and its value is surely dependent on the mutation rate per nucleotide, as well as on ability of the poly-nucleotide strand to survive for subsequent replication. This length also defines the maximum amount of the information content in the early RNA-protospecies, which could have replicated without help of the RNA-replicase ribozyme.

The goal of this study is estimation of the complexity threshold in the early phase of RNA-World, after the stage of very short RNA oligonucleotides (up to 20-30 units). The latter phase, extensively simulated by Ma et al. (2007a) and Myszor and Cyran (2010) (see also section 6.4), is characterized by the non-negligible probability of restoring the sequence of oligonucleotide strand from scratch, i.e. for oligos of the length λ being considerably less than 50. The value 50 has been chosen and the symbolic boundary of phases, since the

probability of restoring from scratch the sequence of 50 nucleotides is equal $4^{-50} = 2^{-100}$ i.e. it is smaller than 10^{-30}, thus it can be safely considered as negligible.

Therefore, it can be safely concluded that the sequences composed of more than 100 nucleotides must have occurred in the continuous evolution of shorter sequences rather than by *ab-initio* creation. The consequence of this fact is that once a given lineage of sequences (proto-species) becomes extinct it practically cannot be brought back to existence (unless highly improbable process of random setting of required nucleotides would have happen).

With the aforementioned assumptions, consider the RNA-species with the RNA chain of the length λ, where $\lambda > 50$. Let such species replicate with the mutation rate per nucleotide equal μ. Then, the probability that the single nucleotide in a strand is copied without an error is given by $p = 1 - \mu$. This yields in a model of independent nucleotides replications the probability of correct replication of the whole polynucleotide strand equal $v = p^{\lambda}$.

Consider also three situations designated by S_0, S_1 and S_2 yielding in the next generation 0, 1, and 2 individuals respectively. Denoting the probability that RNA strand is not hydrolyzed by w, it is obvious that situation S_0 takes place when, with the probability $1-w$, individual does not survive to replication stage (next generation) because of hydrolysis. Similarly, S_1 denotes the situation when, with the probability $w(1- v)$, the individual is not hydrolyzed at least until next generation, but it produces a copy of itself with an error. Finally, situation S_2 denotes the case when, with the probability wv, the individual not only is not hydrolyzed but also it replicates without error yielding two identical strands.

In the further analysis it is assumed that the Demetrius/Kimmel model is used. It proposes that the population of error-free RNA strands follows the Galton-Watson branching process (see Fig. 3.6:1) with the number of individuals Z_t at time t, given by (3.6:1), (3.6:2), or equivalently (3.6:10). As said in section 3.6 based on formula (3.6:17), such a process is said to be supercritical when the probability of its eventual extinction q satisfies inequality $q < 1$. This happens only when $E(X) > 1$, where random variable X (denoting the number of descendants of given individual chain) is equal zero, one or two with probabilities of situations S_0, S_1, and S_2 respectively. Interestingly, based on formula (3.6:17), even when $\lim_{t\to\infty} E(Z_t) = 1$ for $E(X) = 1$, the probability of eventual extinction $q = 1$ in this critical case, despite the result looks somewhat counterintuitive.

Consider now the probability generating function $f(s)$ of the progeny number in the branching process modeling the evolution of the early RNA-protospecies. It follows that $f(s)$ is given by

$$f(s) = 1 - w + w(1-v)s + wvs^2. \tag{6.2:1}$$

Therefore, the probability of extinction q, being the smallest positive root of the equation $f(s) - s = 0$, with roots $q_1 = 1$ and $q_2 = (1 - w) / wv$, is obviously equal to q_2. Of course in

order to avoid the necessity of extinction (with probability one), q_1 must be greater than q_2, which yields the inequality

$$\frac{1-w}{w} < v. \tag{6.2:2}$$

Even if inequality (2) is satisfied there is a chance of extinction, which can happen with probability $P(ext)=q_2$ and long-term survival of species is expected only with probability $P(surv)$ given by

$$P(surv) = 1 - \frac{1-w}{wv}. \tag{6.2:3}$$

Result described by (2) can be obtained also directly from the criticality condition expressed as the inequality $E(X) = f'(1) = w\,(1+v) > 1$. Indeed, the last statement is satisfied only when formula $v > (1 - w) / w$ given by inequality (2) holds.

Substituting $v = (1 - \mu)^{\lambda}$ and solving with respect to λ there is obtained the complexity threshold satisfying

$$\lambda < \frac{\ln(1-w) - \ln w}{\ln(1-\mu)}. \tag{6.2:4}$$

The above formula does not take into consideration the fact that probability w of avoiding by RNA strand the hydrolysis at least to the subsequent replication event is also dependent on the length of the strand λ.

To introduce this dependency, consider more detailed model in which parameter r denotes the probability of breaking the phosphodiester bond between nucleotides in the RNA strand in the time between successive replications. Since in a strand of λ nucleotides there are $\lambda - 1$ bonds, therefore $w = (1 - r)^{\lambda - 1}$. In the new model it is impossible to obtain explicit formula for λ so feasible values should be computed numerically from inequality

$$1 < (1-r)^{\lambda-1}\left(1+(1-\mu)^{\lambda}\right). \tag{6.2:5}$$

Assuming that the complexity threshold, denoted as $\lambda_{critical}$, is defined as such λ for which formula (5) modified to be an equation holds, the critical mutation rate $\mu_{critical}$ is given by

$$\mu_{critical} = 1 - \left(\frac{1}{(1-r)^{\lambda_{critical}-1}} - 1\right)^{\frac{1}{\lambda_{critical}}}. \tag{6.2:6}$$

For all mutation rates larger than $\mu_{critical}$ RNA species become extinct with probability one. In Fig. 1, the 3D plot of the border function for $\mu_{critical}$ is presented for range of parameter $\lambda_{critical}$ from 1 to 10^3 and range of parameter r from 10^{-4} to 10^{-3}. Fig. 2 presents similar plot for parameter r ranging from 10^{-5} to 10^{-4}.

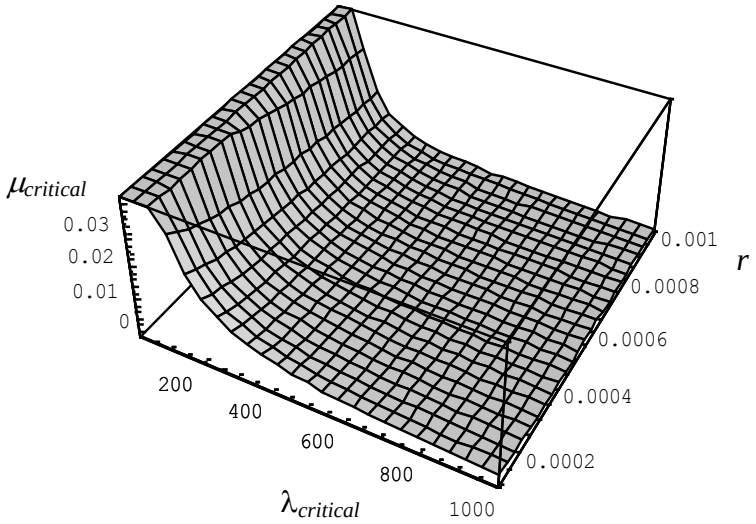


Fig. 6.2:1. Surface of the function $\mu_{critical}$ ($\lambda_{critical}$, r) for r ranging from 10^{-4} to 10^{-3} and $\lambda_{critical}$ ranging from 1 to 10^3

Rys. 6.2:1. Powierzchnia funkcji $\mu_{critical}$ ($\lambda_{critical}$, r) dla r z zakresu od 10^{-4} do 10^{-3} i $\lambda_{critical}$ z zakresu od 1 do 10^3

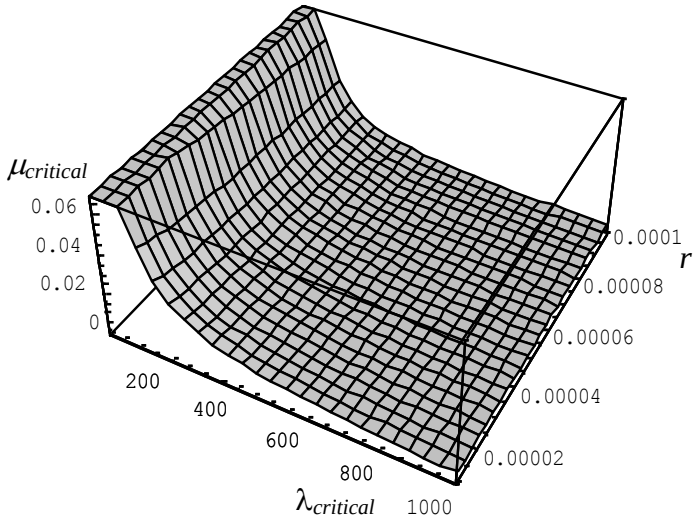


Fig. 6.2:2. Surface of the function $\mu_{critical}$ ($\lambda_{critical}$, r) for r ranging from 10^{-5} to 10^{-4} and $\lambda_{critical}$ ranging from 1 to 10^3

Rys. 6.2:2. Powierzchnia funkcji $\mu_{critical}$ ($\lambda_{critical}$, r) dla r z zakresu od 10^{-5} do 10^{-4} i $\lambda_{critical}$ z zakresu od 1 do 10^3

It is clearly visible that for larger values of $\lambda_{critical}$ the critical mutation rate $\mu_{critical}$ must be smaller. Not surprisingly, the slope of this function decreasing with $\lambda_{critical}$ is steeper for higher probabilities of the phosphodiester bond break r. Since in all experiments of the evolution of RNA strands in abiotic conditions in a test tube, the researchers have yielded

mutation rates greater than 10^{-2}, this value can be treated as the cutoff level for the surfaces presented in Fig. 1 and 2. Only points with such coordinates ($\lambda_{critical}$, r) for which the surface of the function $\mu_{critical}$ is above this cutoff represent conditions feasible for long-lasting evolution of the RNA protospecies, which avoid the error catastrophe.

The surfaces presented in Fig. 1 and 2 provide a lot of qualitative information about the character of function representing the critical mutation rate with respect to the complexity threshold and the probability of the break of a phosphodiester bond. One of the most interesting features, which can be studied from these figures, is the monotonicity of the two-dimensional function. However, it is impossible to read from these charts the quantitative characteristics. Therefore, having in mind the monotonic course of the function $\mu_{critical}$ with respect to r, instead of presenting two-dimensional surfaces, Fig. 3, 4, and 5 show one-dimensional curves for values of parameter r equal to 10^{-3}, 10^{-4}, and 10^{-5} respectively. Such three values of this parameter are representative for the wide range of this parameter varying from 10^{-5} to 10^{-3} because of monotonic character of the function with respect to this parameter. The limits for this range have been chosen basing on Ma et al. (2007a), and they will be discussed further in Conclusions (section 6.5). Additionally, for better illustration, the mutation rate cutoff value 10^{-2} was subtracted from the function $\mu_{critical}$, defined by (6), so the resulting plots cross the horizontal axis exactly at the point indicating the complexity threshold $\lambda_{critical}$.

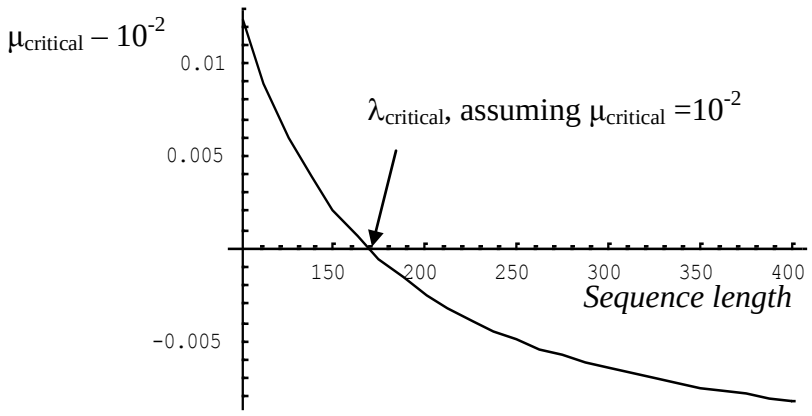


Fig. 6.2:3. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-3}$
Rys. 6.2:3. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-3}$

While in Figures 3, 4, and 5 it is assumed that the mutation rate per nucleotide $\mu_{criticail} = 10^{-2}$, such rate is reported to be a limit of the accuracy in replication without help of the RNA replicase, rather than the actual accuracy. Compare for example review performed by Dyson (1999): ”*All the experiments that have been done with RNA replication under*

abiotic conditions give error rates of the order of 10-2 at best". The most of experiments yield this rate to be as big as 2×10^{-2} or even 5×10^{-2}, as it is reported by Smith and Szathmary (1999), who, what is worth to be stressed, are the advocates of the RNA-world: "*The error rate depends on the medium, the temperature, and so on, but very roughly the wrong base pairs with a G once in 20 times*".

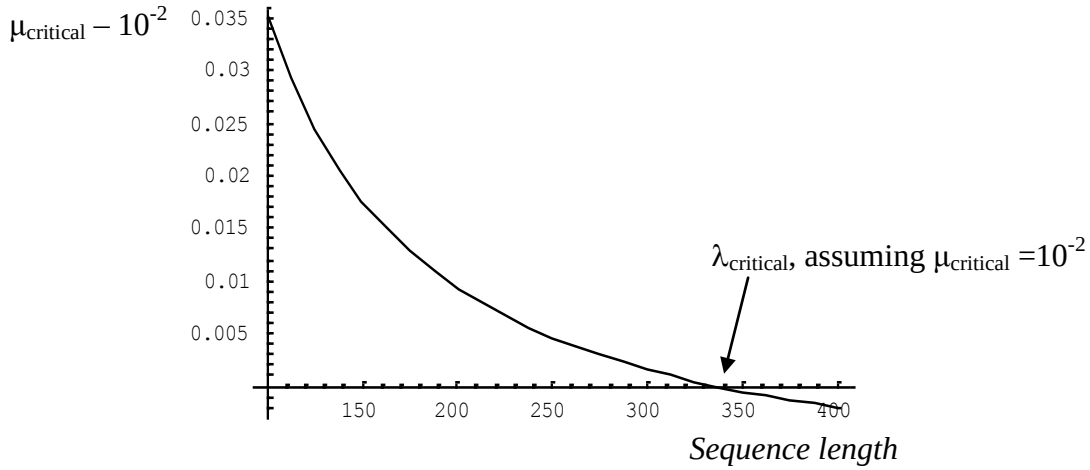


Fig. 6.2:4. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-4}$
Rys. 6.2:4. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-4}$

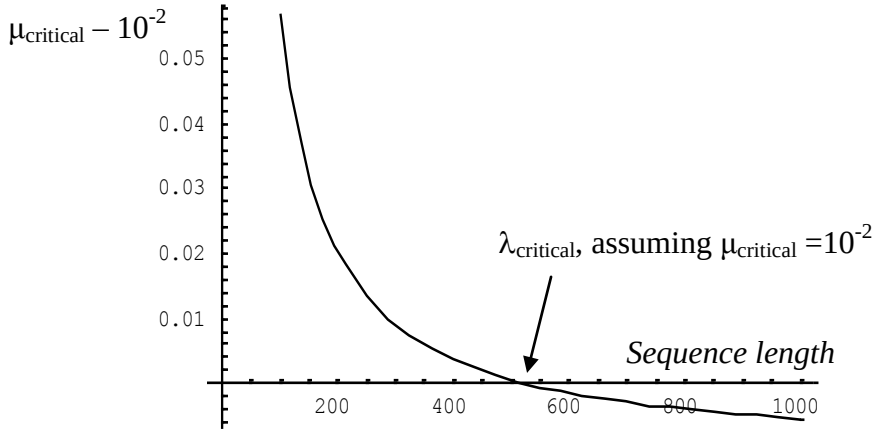


Fig. 6.2:5. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-5}$
Rys. 6.2:5. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-5}$

Having this in mind, below there is presented a set of Figures 6, 7, and 8, as well as a set of Figures 9, 10, and 11, which are the counterparts of a set of Figures 3, 4, and 5 where the mutation rates per nucleotide $\mu_{criticail}$ are 2×10^{-2} and 5×10^{-2} respectively.

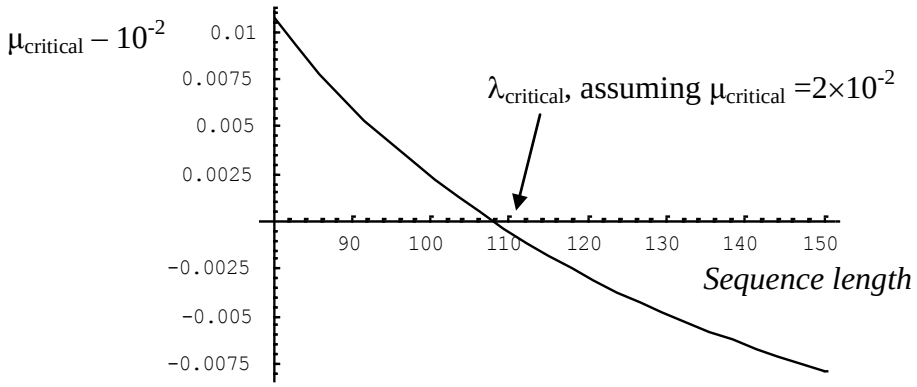


Fig. 6.2:6. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-3}$
Rys. 6.2:6. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-3}$

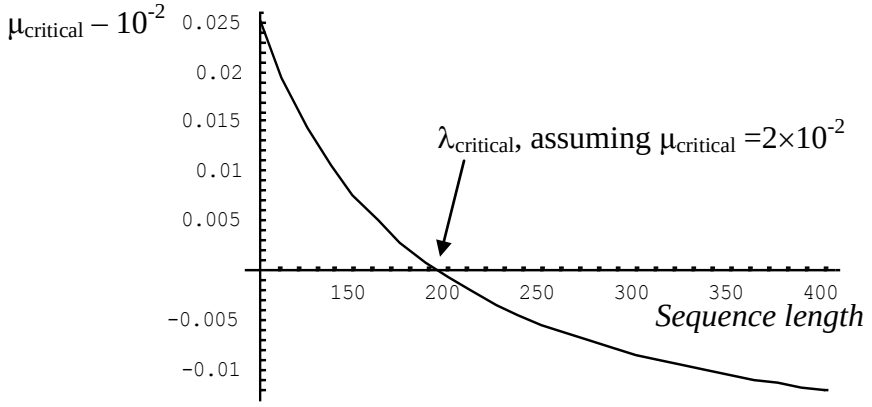


Fig. 6.2:7. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-4}$
Rys. 6.2:7. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-4}$

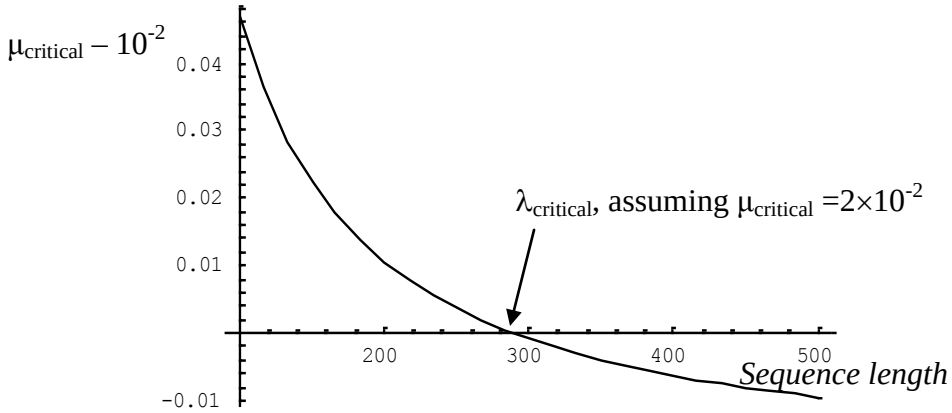


Fig. 6.2:8. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-5}$
Rys. 6.2:8. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-5}$

The problem with RNA-world is that the ribozyme, crucial for replication of any RNA-species, called RNA-replicase is yet to be discovered. Many advocates of the RNA-world do

not consider this lack seriously, believing that it is only the matter of time when experimental confirmation happens. This belief is based on strong foundations, since some enzymatic activity exhibited by RNA molecules has been already demonstrated. Extending the range of discovered ribozymes to RNA-replicase is only one step further.

Perhaps the above line of argument is correct, however, except for the only hypothetical existence of RNA-based RNA-replicase, there exists at least one more serious problem and the study focuses on it. This problem is caused by the possible error catastrophe which can easily occur when RNA strands try to replicate in abiotic conditions. All experiments of the evolution of RNA performed in a test tube indicate that without help of replicase enzyme the error of replication is larger than 10^{-2}. Even if we assume that instead of protein-based enzyme, the ribozyme could be used in the RNA-world, there had to be a period when even primordial replicase had not yet evolved.

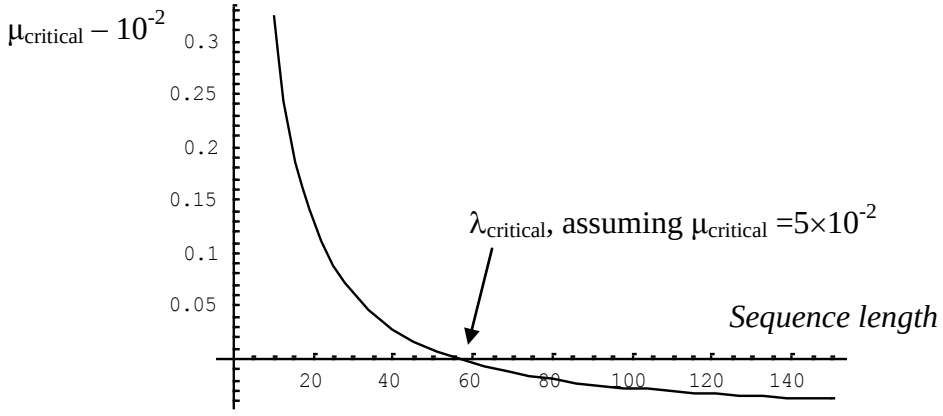


Fig. 6.2:9. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-3}$
Rys. 6.2:9. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-3}$

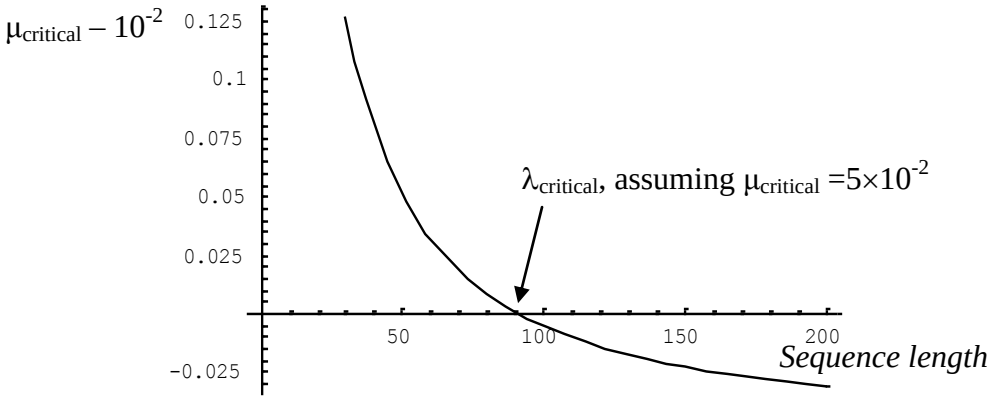


Fig. 6.2:10. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-4}$
Rys. 6.2:10. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-4}$

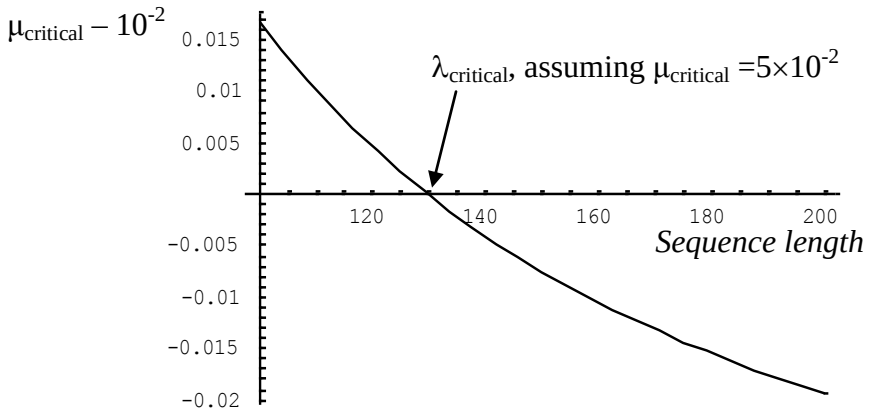


Fig. 6.2:11. The complexity threshold (at intersection of the curie with horizontal axis) for $r = 10^{-5}$
Rys. 6.2:11. Granica złożoności (na przecięciu krzywej z osią poziomą) dla $r = 10^{-5}$

The results, which have been obtained for the complexity threshold, have direct influence on the amount of information which can be preserved in a population of evolving strands (precursors of genes). The problem of the number of genes, which can be replicated and randomly assorted in a compartment model, is the subject discussed below.

6.3. Compartment model with random assortment of genes

In this section the compartment (or package) model of the early life is considered. Compartment model was created as an alternative to the hypercycles model. In the hypercycle model every gene is responsible for encoding polypeptide supporting replication of the next gene in a cycle (Eigen and Schuster 1977). Since the creation of this theory researchers have been arguing about stability and possibility of surviving of such units. Package model proposed by Niesert et al. (1981) is an alternative in which one gene, the primordial replicase, is responsible for replication of all genes in a protocell and the assortment of genes during reproduction is performed in a random fashion.

At the beginning, this primitive replicase couldn't achieve high fidelity level because of RNA strand length constraints discussed in section 6.2. The circle seems to be closed: those limitations in the length are being caused by high mutation level which is caused by low fidelity of replicase. However, with time, and paradoxically thanks to series of luckily mutations, fidelity of replicase should improve.

In her next study, Niesert (1987) used Univac 1100 to investigate model properties. In particular, she was interested in determining the maximal amount of different types of genes (MDTOG) in a package under different mutation rates and different number of replicating molecules (NORM). Computers at that time were slow and the execution time was pretty expensive, so there was a necessity to limit the amount of simulations. Currently, there is a

broad access to computers with 4 core processors inside. In the study, ten devices with Intel Core 2 Quad 2.8 GHz were used. Those powerful machines let us conduct enough amount of simulations to speak about problem in the language of statistics.

In the package model considered, the primordial genes are enclosed in primitive compartments – protocells. A set of protocells is denoted as population. Compartments contain many different types of genes. In one protocell there might be many copies of the same type of gene. In order to survive package must have at least one representative of every type of gene. All genes have an equal replication rate. Once in a while, the package is being split into the two daughter packages. This process is being called as package fission. During the fission all genes from the mother package are being distributed randomly between the daughters packages. If the daughter package doesn't have representative of all gene types it is being dismissed.

The model tries to answer the question: what is the maximum number of different types of genes (MDTOG) in a compartment, which does not lead to the population extinction (caused for example by lacking of one or more types of genes in compartments). Genes are being replicated between the package fissions. The number of replicated genes between package fission is denoted as NORM. This is very important parameter of the model, as intuitively if the number of replicating genes is small there is relatively large probability that in the daughter cell some gene type will not be present. However, if mutation is taken into account, then too large NORM will increase the probability of the lethal mutation, what leads to the extinction of the lineage. Therefore there exists some optimal value for NORM, for which the MDTOG is maximized.

Since it is hard to imagine that NORM could be a constant in such primitive cells, hence the NORM the variation of this parameter was applied in simulations as described below. Genes are being replicated with some fidelity, and as a result, during replication an error can take place, when the mutation occurs. In the model there are distinguished two types of mutation, parasite mutation and the lethal mutation.

Parasite mutation leads to disability of the gene functionality, however, the new gene is not harmful for the package. The only negative impact on the package survival is caused by the fact that such gene might be replicated so it reduces the amount of health genes replicas in a package. Parasite can never become functional gene again. Lethal mutation leads to creation of a gene with disabled functionality which cause an instant death of the protocell or which have remarkably higher replication rate than other genes in a package, what will eliminate descendants of the package in several generations.

In the model only harmful mutations are considered - the mutations which can lead to an improvement in the gene functionality is not taken into account. Lengths of genes are varied by the modification of the value of mutation rate per gene. Additionally, the package can

become a victim of a harmful event which leads to its death. Such an event is called an accident and it is determined instantly after package creation. Mentioned processes are connected with the following parameters of the model: parasite mutation rate (PMR), lethal mutation rate (LMR) and the accident rate (AR).

In the study, at first, the Niesert's simulations (Niesert et al. 1981) were repeated with constant NORM value and different levels of mutations PMR and LMR. Then, the more advanced model (Niesert 1987), which included NORM variation at the package level, was considered. Finally, the new type of NORM variation was introduced by Myszor and Cyran (2009), which was responsible to reflect changes in the environment. In the most advanced model the NORM variation at the package level with NORM variation at the environmental level was considered.

Let us denote by a scenario the simulation with given mutation rate and NORM variation. For each scenario there were created 100 unlinked histories and each history was simulated independently for 1000 succeeded generations. The maximum size of the population was constant and equal 25 packages. In foster conditions the amount of packages can raise exponentially, what can eat up whole computer memory. That is why a limit for 25 packages was established. If there were more packages after creation of the new generation, the reduction of the amount of packages followed. For each package the prospective value proposed by Niesert et al. (1981) was computed and the weakest packages were disposed to keep the limit of the population size.

It was assumed that package is viable only if it possesses at least one copy of each type of genes and at the same time it does not have any lethal gene. The simulated history was considered successful if the last generation possesses at least one viable package. In order to succeed (to be approved) the scenario must have at least five successful histories out of 100, if the goal is to show that the null hypothesis stating that the scenario is not feasible, should be rejected as significance level 0.05. Five survived populations out of 100 might seem to be a small fraction, but truly, in order to create variety of nowadays life one such history was enough. However, in Fig 1, not only the conditions for survival of 5% of histories are presented, but also the conditions for survival of 95% of histories are plotted, for comparison, how the conditions influence the fraction of survived histories.

Fig. 1 presents the situation when the mutation and accident rates are equal zero (black dots and gray crosses) and when the parameters have the following values: PMR = 0.1, LMR = 0.01, AR = 0.01 (gray dots and black crosses). In Fig. 1 the scenario needs 5 (gray dots and gray crosses) or 95 (black dots and black crosses) successful histories in order to be considered as successful.

The differences between 5% and 95% of successful histories to approve scenario for case with turned-off mutations and the accident rate set to 0 and fixed NORM, might seem small,

but when the mutations and the accident rate is turned on, then this difference is more visible (Fig. 1). In this latter case, it influences the value of MDTOG, which is 4 (gray dots for NORM between 25 and 35) when it serves to reject the null hypothesis that the compartments with four types of genes cannot survive, or 3 (black crosses for NORM between 5 and 55) when the goal is to show that population composed of compartments with three types of genes survives in 95% of simulated histories.

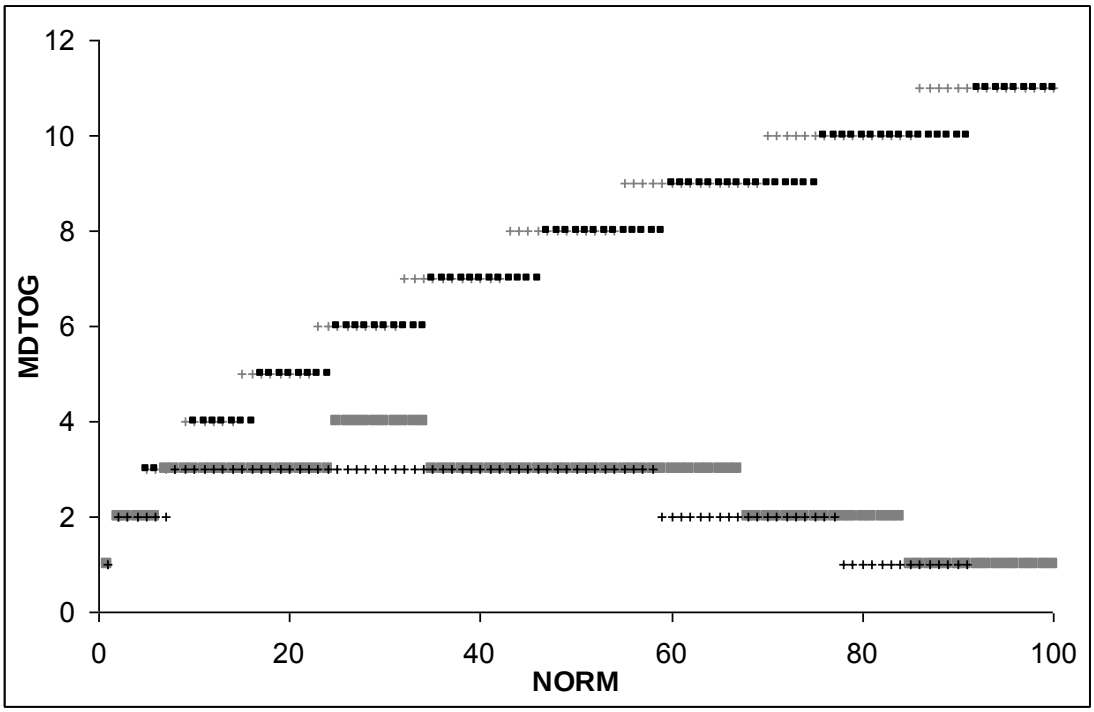


Fig. 6.3:1. MDTOG as a function of NORM
Rys. 6.3:1. MDTOG jako funkcja NORM

As it was stated, two types of variations - at the package level (representing the individual's diversity) and at the population level (representing the changing environment) were studied. The variation at the package level, which was proposed by Niesert (1987) was claimed by her to have small influence on MDTOG. In the study of Myszor and Cyran (2009) simulations with different distribution of NORM have been performed. For this purpose, the normal distribution was used with mean set to the base NORM value and with standard deviation equal to 15 or 30. In Fig. 2, there are presented the results for PMR = 0, LMR = 0, AR = 0, and the normal distribution of NORM. The mean of the distribution is set to the value indicated as the Base NORM at horizontal axis (this convention is used also in subsequent figures), and standard deviation set to 15 (gray dots) and 30 (black x). For comparison, the switched-off variance is indicated by crosses.

The results (Fig. 2) show that there are indeed small differences when mutations and accidents are turned off as compared to results presented in Fig. 1, when no NORM variation has been applied. More significant differences could be observed in Fig. 3, which reports the results for mutations rates and accident rate are different than zero (PMR = 0.1, LMR = 0.01, AR = 0.01, standard deviation set to 15 (black dots) and 30 (gray x), for crosses the variance is off). In this latter case the NORM variation at the package level might significantly reduce the MDTOG, and greater variation results in greater MDTOG reduction.

The variation of NORM parameter at the population level, is new in the model. This variation has been introduced to check whether fluctuating environmental conditions represented by changes in the amount of replicated genes between generations, might have influence on the MDTOG value. For normal distribution of the NORM, there was two cases considered: a) the mean was set to the base NORM and the standard deviation was fixed and set to the defined value (Fig. 4, Fig. 5), and b) the mean was set to the base NORM and the standard deviation was the ratio of the base NORM and some constant (Fig. 6, Fig. 7).

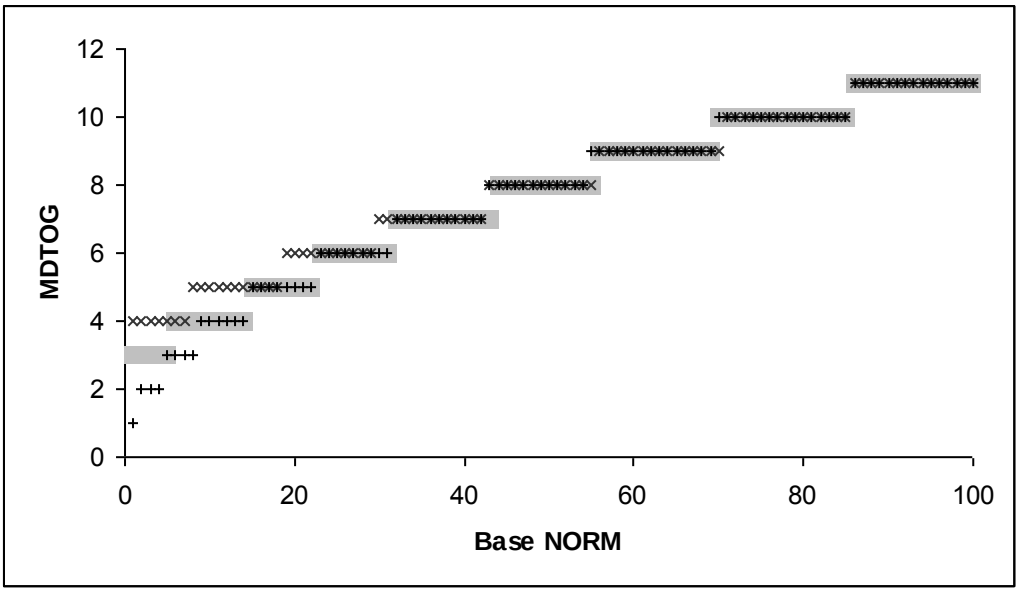


Fig. 6.3:2. MDTOG as a function of expected value of normally distributed within package NORM. PMR = 0, LMR = 0, AR = 0
Rys. 6.3:2. MDTOG jako funkcja wartości oczekiwanej zmiennej NORM o normalnym rozkładzie zmienności wewnątrz kompartmentu. PMR = 0, LMR = 0, AR = 0

In Fig. 4 and Fig. 5 standard deviation is set to 15 (gray x) and 30 (gray crosses), and the variation is switched off for black dots, whereas in Fig. 6 and Fig. 7 standard deviation is proportional to the Base NORM, according to: BaseNorm/5 (gray x), and BaseNorm/2 (gray crosses). For black dots the variation of the NORM parameter is switched off. Presented

graphs imply that environmental impact is significant, especially for fixed standard deviation. It follows that the variation of NORM across population level reduces the MDTOG.

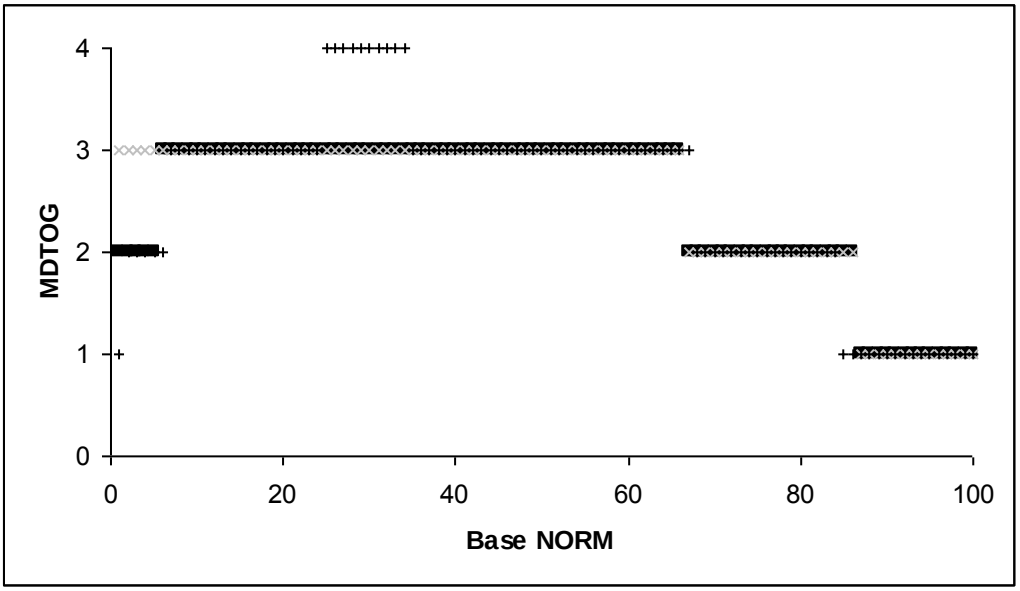


Fig. 6.3:3. MDTOG as a function of expected value of normally distributed within package NORM. PMR = 0.1, LMR = 0.01, AR = 0.01
Rys. 6.3:3. MDTOG jako funkcja wartości oczekiwanej zmiennej NORM o normalnym rozkładzie zmienności wewnątrz kompartmentu. PMR = 0.1, LMR = 0.01, AR = 0.01

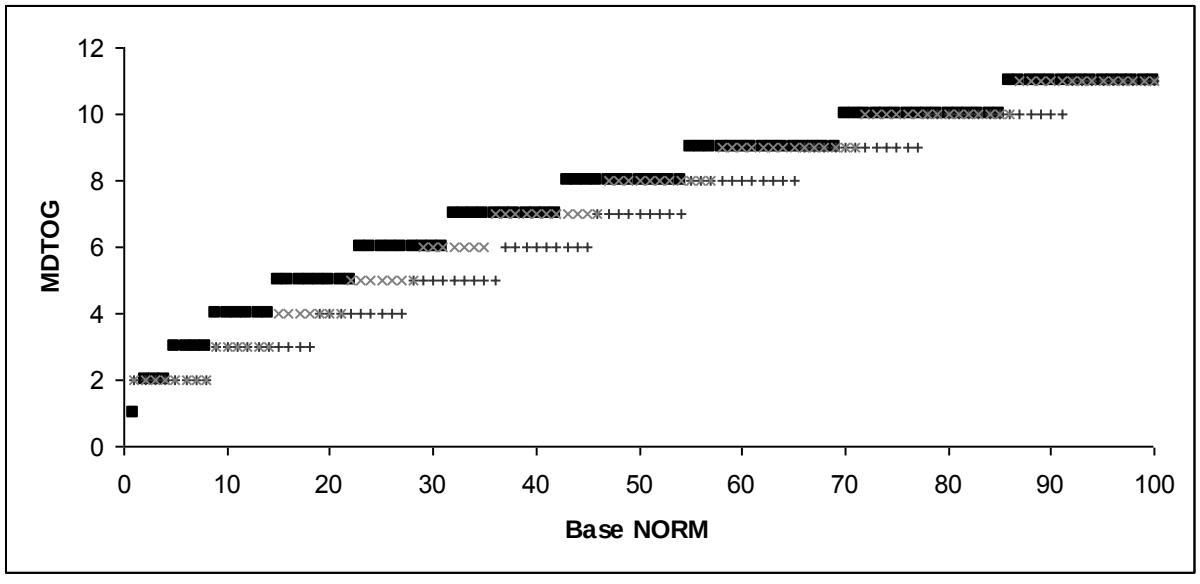


Fig. 6.3:4. MDTOG as a function of expected value of normally distributed across populations NORM (constant variance). PMR = 0, LMR = 0, AR = 0
Rys. 6.3:4. MDTOG jako funkcja wartości oczekiwanej zmiennej NORM o normalnym rozkładzie zmienności w populacji. PMR = 0, LMR = 0, AR = 0

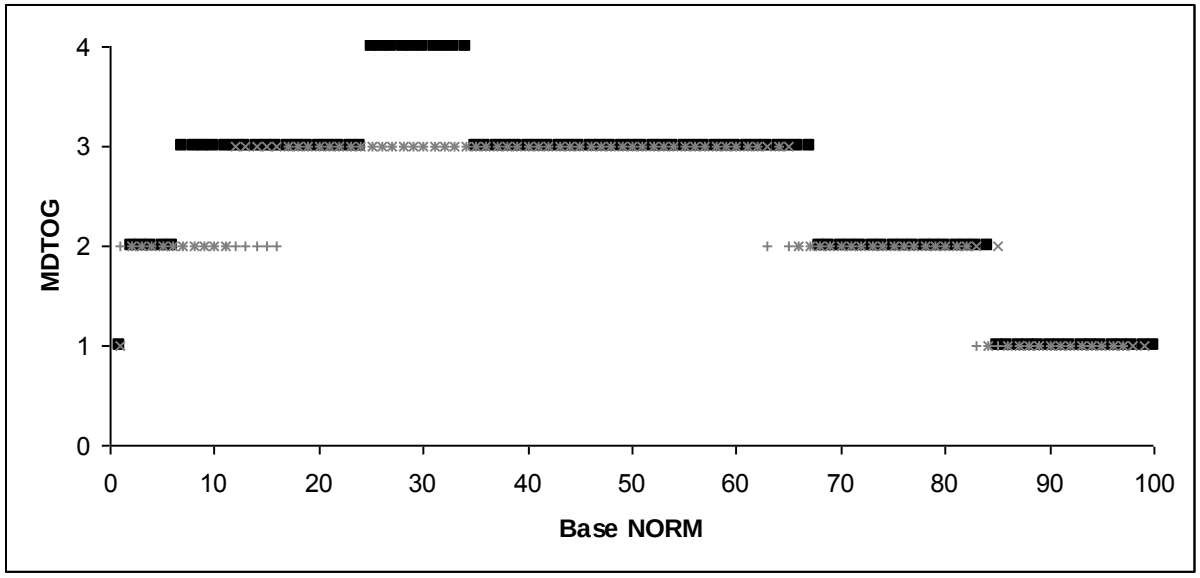


Fig. 6.3:5. MDTOG as a function of expected value of normally distributed across populations NORM (constant variance). PMR = 0.1, LMR = 0.01, AR = 0.01
Rys. 6.3:5. MDTOG jako funkcja wartości oczekiwanej zmiennej NORM o normalnym rozkładzie zmienności w populacji. PMR = 0.1, LMR = 0.01, AR = 0.01

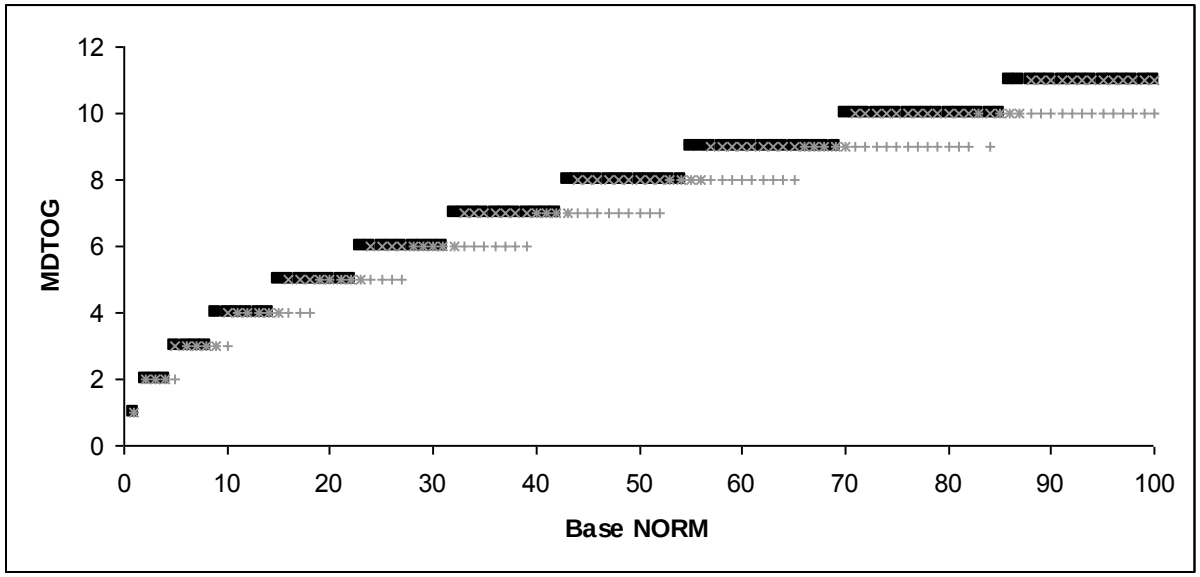


Fig. 6.3:6. MDTOG as a function of expected value of normally distributed across populations NORM (variance proportional to NORM). PMR = 0, LMR = 0, AR = 0
Rys. 6.3:6. MDTOG jako funkcja wartości oczekiwanej zmiennej NORM o normalnym rozkładzie zmienności w populacji (wariancja proporcjonalna do NORM). PMR = 0, LMR = 0, AR = 0

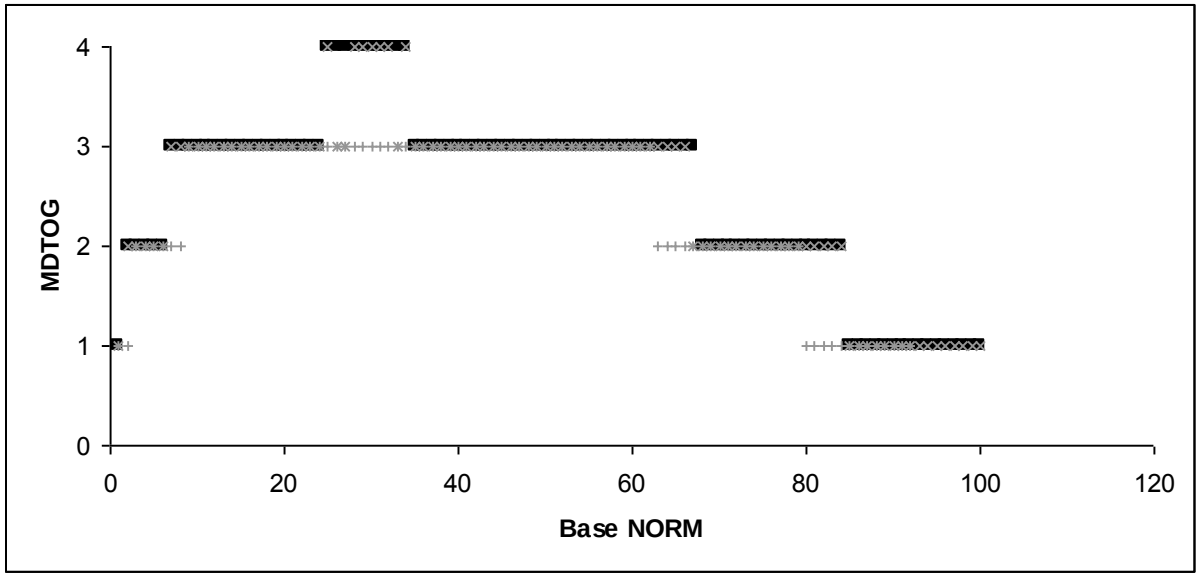


Fig. 6.3:7. MDTOG as a function of expected value of normally distributed across populations NORM (variance proportional to NORM). PMR = 0.1, LMR = 0.01, AR = 0.01
Rys. 6.3:7. MDTOG jako funkcja wartości oczekiwanej zmiennej NORM o normalnym rozkładzie zmienności w populacji (wariancja proporcjonalna do NORM). PMR = 0.1, LMR = 0.01, AR = 0.01

After a series of introductory experiments, the goal of the study was to check whether the packages could exist without comprising the replicase as a one of the genes in a compartment. For that purpose, simulations were performed with NORM within-package variation (standard deviation set to 15 or 30). The experiment showed that without replicase, i.e., when PMR is estimated to be 0.01 per nucleotide (which is the most optimistic result of all experiments with evolution in a test tube – see section 5.2 for details), and further optimistically assuming that LMR and AR are both equal zero, the compartment could have maximally two types of genes, each 50 nucleotides long (such result yields mutation per gene equal 0.39) or one type of gene containing 100 nucleotides (mutation equal 0.63 per gene). For greater mutation rate, such as equal 0.02 per nucleotide (what seems to be more realistic value for replication without replicase), there might be maximally one type of gene in a package, even if it has only 50 nucleotides.

The simulations also showed that there is a limit in the length of a single gene in a package even for PMR = 0.01 per nucleotide. In such a case, this limit is close to 500 nucleotides (PMR = 0.99 per gene, with LMR = 0, AR = 0). For more real conditions with LMR set to 0.005 and AR set to 0.01, the maximal length of the gene is 200 nucleotides. Such amount of information in the package is similar to the amount of information capable to be preserved in a single strand model. Therefore, those results are coherent with author's studies (Cyran 2009b) concerning complexity threshold in the single strand model described in section 5.2.

Above results clearly indicate that the only advantage of the compartment model in terms of the amount of information stored in the genome is when one of the genes is the replicase. It can increase the replication fidelity ten-fold so the mutation rate in the best case might be equal to 0.001 per nucleotide. If the typical primordial gene in a package would have 100 nucleotides, then the mutation rate per gene would be close to 0.01 and for 200 nucleotides close to 0.02. A series of simulations were conducted for such values with environmental variation and without it, and with different LMR and AR values.

For typical conditions (LMR = 0.01, AR = 0.01 and PRM = 0.01), MDTOG was equal to 4 without NORM variation, and 3 for environmental NORM variation. When PRM was set to 0.02 the results are even more pessimistic: MDTOG is equal 2 without and with environmental variation. Moreover, in this latter case, the package might contain 2 different types of genes within very narrow NORM range. Taking in mind that one of these genes has to be replicase, there is a room for only one additional gene in a compartment with randomly segregation of genetic material. Thus, the next step in evolution, had to be the "invention" of chromosomes, which assure linking of genes, and hence, non-random assortment.

6.4. Non-enzymatic template-directed RNA recombination model

Since many years scientists have wondered how life emerged. There were many trials of explanation of this process, however up to now they are non-fully succeeded. Currently, researches shed new light on some chemical processes that might play important role during origin of life and should be taken into account in studies concerning origins' reconstruction. The currently most popular and consistent – despite problems raised by Dyson (1999) – theory, which describes life's beginning is, no doubt, the RNA world (Orgel 2004). According to this hypothesis there was the time when life based on RNA strands, and these strands could store information and act as chemical reaction catalysers (Joyce 2005, Cochrane and Strobel 2008, Steitz and Moore 2003).

There are many proofs of this theory visible in current life, for example ribozymes (Joyce and Orgel 2006). However, there are still many pieces of the puzzle that just do not fit. One of such unfitted pieces is the problem with the RNA strands lengths. According to the RNA world hypothesis, nucleotides, after emergence in the solution, might join to each other and form strands. The binding of the RNA molecules might be the effect of a process of the mineral-catalyzed synthesis of polynucleotides (Ferris and Ertem 1993, Ferris et al. 1996, Ferris 2002, 2006). Outcomes of laboratory experiments indicate that there is a possibility to acquire chains up to 50 nucleotides long (Huang and Ferris 2006). Scientists speculate that further elongation of RNA chains might be acquired through RNA recombination processes

(Lutay et al. 2007). There are three types of RNA recombination: non-enzymatic, non-enzymatic template-directed, and RNA-directed.

Non-enzymatic recombination is a process in which two RNA chains join together with complementary nucleotides. In the next step each strand looses superfluous part in the cleavage reaction and then, through ligation reaction, strands became connected by phosphodiester bond (Lutay et al. 2007). Fig. 1 presents this process in three following steps: polynucleotide approach (left), complementary parts attraction (center), and cleavage and ligation reaction (right).

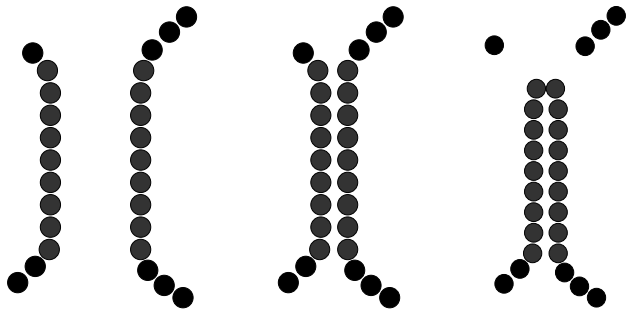


Fig. 6.4:1. Two polynucleotides strands with complementary nucleotides, connected in non-enzymatic recombination process
Rys. 6.4:1. Dwa łańcuchy polinukleotydowe z komplementarnymi nukleotydami połączone w procesie nieenzymatycznej rekombinacji

In the second type of recombination, the non-enzymatic template-directed one, there are short RNA strands in the solution, which serve as templates. To these RNA molecules other oligonucleotides might be attached with complementary parts. If two attached strands are located close enough to each other on the template, they might recombine. During the recombination molecules loose superfluous part of chains through cleavage reactions and are joined in ligation reaction.

According to laboratory experiments, connected strands might tightly cling to the template or different formation might be created around the place of a join point, such as 1 or 3 nucleotides bulges, 2-3 internal loops, etc. (Nechaev et al. 2009). This process is shown in Fig. 2, in which two RNA strands (light and dark) attached to the template (black horizontal sequence). Two strands become attached to the template (upper picture), in cleavage reaction loose superfluous parts (middle picture), and finally, they become connected in ligation reaction (lower picture).

The third type of recombination, that which is RNA-directed, assumes the existence of oligonucleotides with a catalytic activity that through binding processes can catalyze direct recombination of other strands in the solution (Draper et al. 2008).

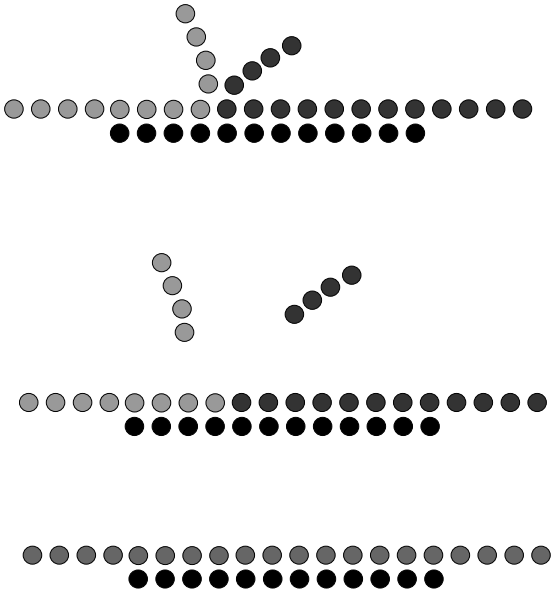


Fig. 6.4:2. Non-enzymatic template-directed recombination process
Rys. 6.4:2. Proces nieenzymatycznej, sterowanej matrycą rekombinacji

Myszor and Cyran (2010) applied computer simulations to check the influence of the non-enzymatic template-directed recombination process on lengths of polynucleotides in the RNA-world. In order to conduct simulations the model proposed by Ma et al. (2007a) was implemented and improved to include more chemical processes (Ma et al. 2006, Ma et al. 2007a, 2007b), in particular the non-enzymatic template-directed recombination (Nechaev 2009). Therefore, it is believed that the simulated model relies more closely on real chemical processes operating on the RNA strands as compared to model proposed by Ma et al. (2007a). Simulated processes take place on a flat, two-dimensional space, divided into rectangular sectors. This is depicted in Fig. 3, where dots indicate constituents, and only constituents in the same sector might react with each other.

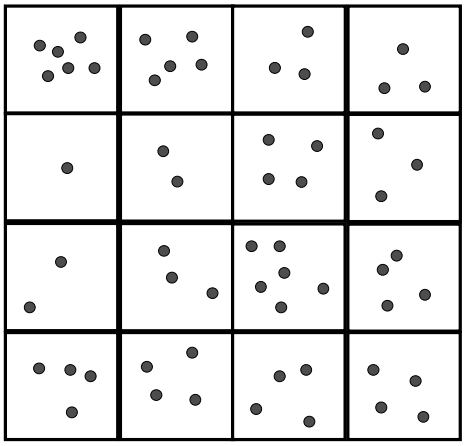


Fig. 6.4:3. Two dimensional surface divided into rectangular sec tors
Rys. 6.4:3. Dwu-wymiarowa powierzchnia podzielona na prostokątne sektory

In the model considered it is assumed, that at the bottom there is a mineral surface that catalyzes polynucleotide formation, and above there is a mixture of chemical substrates. Simulation begins with a set of raw material constituents. Model describes the first phase of the RNA-world before creation of protocells. During the simulation of a single generation, the state of each constituent might be modified only once. The amount of building material in the system is constant (raw material + nucleotides) and it is determined at the beginning of simulation, together with rates of real chemical reactions. All constituents in the model are activated, and the secondary structure of polynucleotides is not taken into account.

There are following types of constituents in the system:

- Raw material: it is a constituent that might become a nucleotide. Recently experiments point out that nucleotides may be created spontaneously from mixture of substrates presented on early Earth, such as 2-aminooxazole, phosphate, part of nucleobases (Szostak 2009). In order to speed up the simulation process there is only one type of raw material constituent in the system.
- Nucleotide: it represents any of four types of nucleotides, cytosine (C), guanine (G), adenine(A), and uracil (U). It might become raw material constituent in a process of degradation. Nucleotide may also join to another nucleotide or polynucleotide in the process of mineral-catalyzed synthesis of polynucleotides.
- Polynucleotide: it is a chain of nucleotides connected by phosphodiester bonds. It might become longer, as a result of mineral-catalyzed synthesis of polynucleotide, attach complementary nucleotides and polynucleotides (become the template), or split into two chains as an effect of phosphodiester bond break.
- Template: it represents a polynucleotide which has a complementary strand attached by hydrogen bonds. It might attach other nucleotides and polynucleotides.
- Attached RNA: it is a RNA chain, which is attached to the template. It might be connected with an adjacent chain by phosphodiester bond in ligation process. It may be de-attached from the template, longer chains are less likely to be de-attached.
- Replicase: this constituent represent a crucial polynucleotide that contains replicase sequence. Replicase sequence is explicitly given at the beginning of the simulation. During simulation, the search is performed for the replicase sequence in each polynucleotide in the system. Replicase itself has got the same properties as other polynucleotides however when bound to other replicase (or strand containing sequence complementary to replicase) it initializes and speeds up the process of complementary strand formation (by the increased probability of molecule attachment and ligation reaction). It also decreases the probability that attached strand drops from the template

before formation of the whole template's copy (Johnston et al. 2001, Zaher and Unrau 2007, Monnard and Szostak 2008).

- Template with bound replicase : it is the RNA chain with the bound replicase polynucleotide. Such template has higher probability of the complementary chains attachment, there is also higher probability of ligation of attached strands. If the whole template with bound replicase has a complementary sequence attached and all the attached nucleotides are joined by phosphodiester bond with adjacent nucleotides then the replicase detaches complementary strand and drops from the template.
- Bound replicase: it represents a replicase, which has bound to a strand containing the replicase sequence or the sequence complementary to replicase. Replicase, which is bound to the template directs attachment of complementary strands and ligation of attached strand. Constituent with bound replicase becomes a template with bound replicase. Replicase might drop from the template before complementary strand formation.

During the simulations the following processes were modeled:

- Nucleotide formation: a raw material might become nucleotide with probability P_{NF}, and a type of nucleobase possessed by formed nucleotide is randomly drawn form A, C, G, U.
- Nucleotide decay: with probability P_{ND} a nucleotide might be broken down into building compounds and become a constituent of raw material.
- Mineral-catalyzed polynucleotide formation: clay, which was common mineral on the early Earth, might act as catalyst of polynucleotides formation. Elongation process uses activated nucleotides and polynucleotides, which are present in the solution. According to the current studies, there is a possibility to create short RNA oligonucleotides up to 50 nucleotides long (Huang and Ferris 2006). The probability of this reaction, P_{MCP}, is given by

$$P_{MCP} = P_{LMC} / L_P \tag{6.4:1}$$

where L_P is the polynucleotide length and P_{LMC} is the probability of ligation by mineral catalysis of two nucleotides. The length of a joined polynucleotide is incorporated into the equation because, in order to be connected, the RNA strands must be aligned in correct way. Longer polynucleotides are less likely to have correct end-to-end orientation. This is presented in Fig. 4, where only polynucleotides that are correctly aligned might react with each other (upper picture) and polynucleotides with incorrect end-to-end alignment are unable to react (lower picture).

- Phosphodiester bond break: degradation process that leads to connection break between nucleotides in the strand. There might be many reasons for phosphodiester bond break

such as: hydrolysis, high temperature, radiation, chemical substances. The phosphodiester bond might be broken with probability P_{BB}.

- Molecule attachment: polynucleotides and templates might attach nucleotides and polynucleotides with probability P_{AT} (or P_{ATR}, for a template with bound replicase). The whole sequence of attached component should be complementary to some part of the template strand (Fig. 5a). There is a possibility of error, and with probability P_{FP}, the nucleotide in the attached sequence might not be complementary to the respective template's nucleotide. Nevertheless, the RNA molecule might still be attached (Fig. 5b).
- Molecule de-attachment: at any time a component might be de-attached from the template with the probability P_{DA} given by

$$P_{DA} = P_{SP} / n, \tag{6.4:2}$$

where n is a number of attached nucleotides and P_{SP} is the probability of the nucleotide separation.

- Ligation: in this process, two adjacent strands attached to the template become connected. Ligation took place with probability P_{LT} for molecules attached to the template and with probability P_{LTR} for molecules attached to the template with the bound replicase.
- Replicase binding: with probability P_{RB} the replicase might bind to the chain containing the replicase sequence or the sequence complementary to the replicase.
- Replicase dropping: replicase might drop from the template during complementary strand formation with probability P_{RD}. Substrate stays on the template, and template with bound replicase becomes template without replicase, and the components acquire default set of probabilities.
- Migration: space is divided into a grid of rectangular sectors, and molecules might migrate to adjacent cell with probability P_M defined as

$$P_M = P_{MN} / w^{1/3}, \tag{6.4:3}$$

where w is the weight of a molecule and P_{MN} is the probability of a move of constituent with a weight equal to 1, such as nucleotide and raw material constituent. The target sector must by adjacent by wall to the current. Moreover, during simulation of a single generation, a component, which should be moved to adjacent cells, is marked,. At the end of simulation of this generation, i.e., after computation of each sector's states, the marked components are moved, accordingly.

Listed above processes are simulated in one of two types of reactions, which are performed in the system. The first type is a reaction, which involves only one constituent, such as nucleotide formation. The second type is a reaction, which involves two molecules, such as polynucleotide formation. During simulation of each generation (one simulation pass)

each constituent is taken from the system and checked whether it should occur in some chemical reaction. If it is true, then depending on the reaction type, the constituent state is modified (for reactions that involve only one constituent) or a search is performed for another molecule to react with (for reactions that involve two constituents).

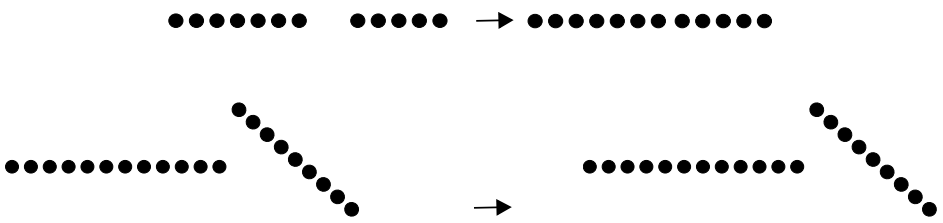


Fig. 6.4:4. Mineral-catalyzed polynucleotide formation
Rys. 6.4:4. Formowanie polinukleotydów katalizowane przez minerały

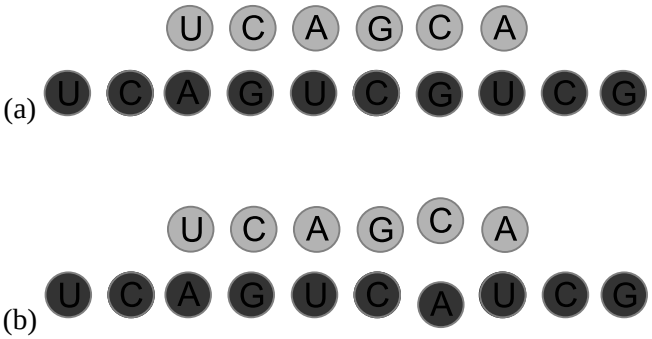


Fig. 6.4:5. (a) Template (black) with attached polynucleotide (gray). (b) Attracted polynucleotide has non-complementary nucleotide C
Rys. 6.4:5. (a) Matryca (czarne) z przyłącznonym polinukleotydem (szare). (b) Przyciągnięty polikunkleotyd ma niekomplementarny nukleotyd C

Since in one simulation pass the constituent's state might be modified only once, hence, the second component, which has been found as a partner for the second type of reactions, is marked as modified and it is not used any more during this simulation pass. Order in which the constituents are checked is connected with constituents' location within sectors. In order to make the model closer to the real world, the constituents are mixed within sectors after each simulation pass.

As a random number generator Mersenne twister (Matsumoto and Nishimuram 1998) was used. It ensures long periodicity and high speed of pseudorandom number generation. Simulation of all sectors demands heavy computations. Fortunately, during the simulation of each generation, there is a time when every sector might be processed independently. In order to achieve results in reasonable time, computation of every sector was executed in parallel, so

when the simulations are run on computer with multi-core processor, then, the whole available computational power is used.

RNA recombination processes are considered as the new hope in the field of the RNA-world theory. There are publications that describe these processes and speculate about possible benefits, however up to this time none of them has incorporated recombination into such exact computer model of the RNA world, as the one considered by Myszor and Cyran (2010). In this latter model, the non-enzymatic template-directed recombination was implemented in extension to reactions simulated by Ma et al (2007a, 2007b).

To some extent, this type of recombination process is similar to the process of non-enzymatic template-directed replication. At the beginning substrates chains are attached to the template. However, contrary to the template-directed replication implemented by Ma et al. (2007a, 2007b), not the whole sequence of the attached chain must be complementary to the template sequence. It is assumed that the sequence might be attached to the template and might recombine with adjacent strand if it has at least four complementary nucleotides in a row with the template sequence.

In order to be able to compare outcomes of these two models, the probability of the component attraction (P_{AT}) and the probability of molecules ligation (P_{LT}) were taken from the template-directed replication and applied also to recombination processes. Additionally, if a nucleotide from a substrate sequence is not matching the respective nucleotide from template sequence it might be accepted with probability P_{FP}. Then, chains located close enough on the template might recombine. During the recombination, the dangling ends of recombining chains are being cut through cleavage reaction and attached strands became connected through the ligation reaction. Strands after the connection might closely adjoin to the template, but there is also a possibility for a creation of more complex structures, in the area of recombined strands conjunction (Nechaev et al. 2009), such as bulge loops of different size on the attached strand, and symmetric or asymmetric loops on the attached strand and on the template (see Fig. 6). In the simulations, the possibilities of creation of the forms presented in Fig. 6 were implemented, what allowed for emergence of such constructions. Probabilities of the creation of these structures are based on real experiments.

Note, that simulation of the recombination have a great impact on a computer model performance. In order to recombine two RNA chains there is a need to find complementary parts of chains with the template. In order to speed-up the search process the suffix-tree algorithm was implemented. It is time-efficient algorithm that made possible to locate common chains' sequences quickly and easily.

At the beginning, the model was simulated without RNA recombination in order to check whether the same results can be obtained as those obtained by Ma et al. (2007a). Then, simulations with non-enzymatic template-directed RNA recombination followed this initial

phase. Finally, the influence of the RNA recombination on the length of polynucleotides was measured using scenario with the replicase strands present in the system.

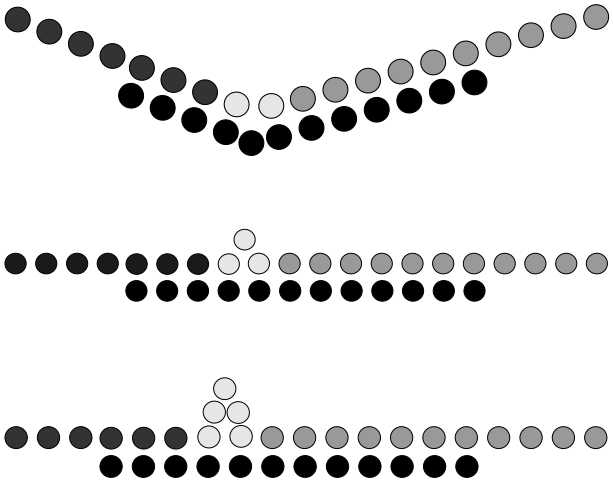


Fig. 6.4:6. Different formations around the place of recombined polynucleotides conjunction after non-enzymatic template-directed RNA recombination process
Rys. 6.4:6. Różne formacje wokół miejsca łączenia rekombinowanych polinuklotydów po nieenzymatycznej sterowanej matrycą rekombinacji RNA

Each simulation started with a set of raw material constituents. In order to achieve trustworthy results, first the simulations were run for 100,000 generations. It was observed that during this period the outcomes were stabilizing. Then simulation process was kept running for subsequent 900,000 generations. The data were collected from each generation dividable by 10,000. In particular, for each chain length, the number of representatives was saved to be used in histograms. After simulation end, the mean number of representatives was computed for each RNA chain length, which occurred during simulation.

Default simulation coefficients were chosen as: grid size 10×10, $P_{NF} = 0.0001$, $P_{ND} = 0.001$, $P_{LMC} = 0.0002$, $P_{BB} = 0.0001$, $P_{AT} = 0.01$, $P_{LT} = 0.005$, $P_{FP} = 0.01$, $P_{SP} = 0.9$, $P_{RB} = 0.95$, $P_{RD} = 0.05$. Note, that the same probabilities notations and values were used as in Ma et al. (2007a). In order to check the influence of the maximal number of constituents in the system, simulations were conducted for different number of constituents in the system, $N = 50{,}000$, and $N = 100{,}000$. For these values, the maximal length of the acquired strands was close to 50 nucleotides, similar to the results obtained in laboratory experiments.

In order to model the non-enzymatic template-directed ligation, the simulations were performed for default probabilities. For $N = 50{,}000$, (Fig. 7) there is a hardy visible reduction of the number of representatives of shorter oligonucleotides (length < 60 nt.).

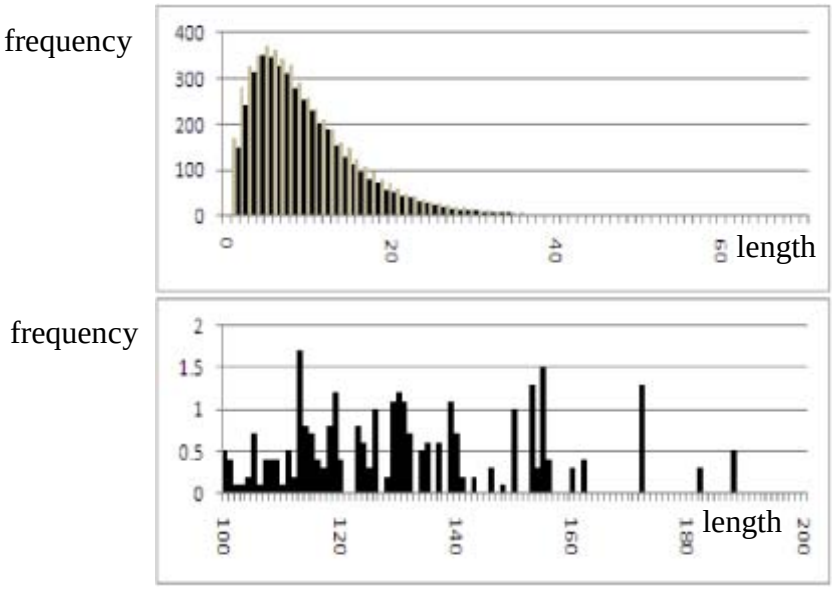


Fig. 6.4:7. RNA molecules lengths without (gray) and with (black) recombination. N = 50,000
Rys. 6.4:7. Długości molekuł RNA bez (szare) oraz z (czarne) rekombinacją. N = 50,000

This reduction is compensated by the fact that significantly longer sequences appear in the solution (length > 100 nt.). It is also worth to mention that in the scenario without RNA recombination such long sequences do not occur. This effect might be hard to notice for scenario with lower number of constituents in the system, however when the number of constituents is increased to N = 100,000 its effect is much larger, i.e., more long strands are then created (Fig. 8).

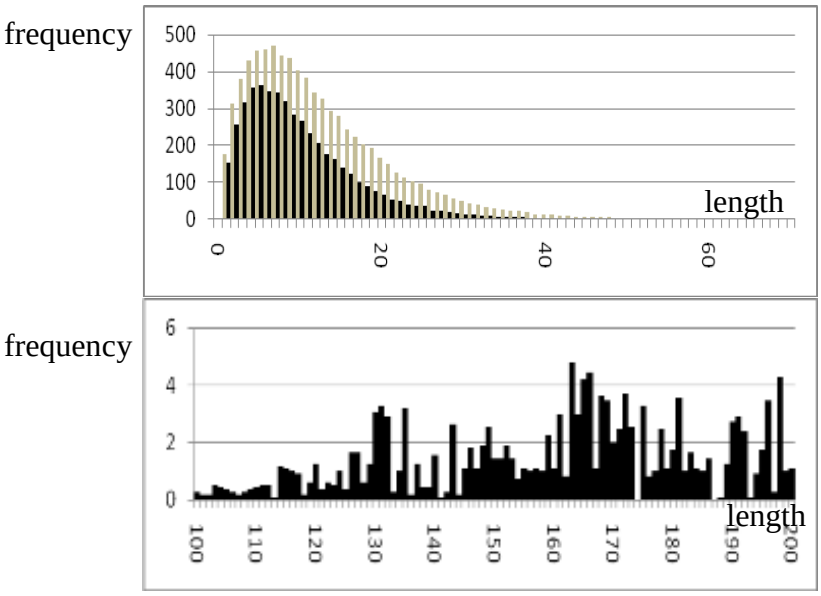


Fig. 6.4:8. RNA molecules lengths without (gray) and with (black) recombination. N = 100,000
Rys. 6.4:8. Długości molekuł RNA bez (szare) oraz z (czarne) rekombinacją. N = 100,000

Interestingly, some current studies suggest that life might emerged in a frozen solution (Vlassov et al. 2004, Kazakov et al. 2006). In a proper temperature the rate of ligation reaction is increasing contrary to the rate of phosphodiester bond breaking. What is more, in lower temperatures fewer intermolecular reaction might be required to stabilize RNA complexes. In order to simulate this phenomena the increased value of P_{AT} was chosen as $P_{AT} = 0.1$ (Fig. 9).

In the next step, the value of P_{LT} was increased to 0.05 (Fig. 10). The results clearly point out that recombination process might be an important phenomena in a frozen solution and leads to the formation of much longer nucleotides than in the model without recombination.

Finally, the influence of the RNA recombination on the replicase emergence was examined as well as the spread of it in the system was modeled. It was assumed that the replication process of template with bound replicase is directed by replicase and is not subject to recombination processes. Simulations were performed for different lengths of replicase sequence. The outcomes point out that the replicase might emerge in such conditions and it might spread in the system (Fig. 11). What is more, the long sequences created by recombination are also present in the solution, however the influence of recombination is limited. This phenomena might be an effect of the limitation of the number of building constituents in the system. Oligonucleotides containing the replicase sequence or the sequence complementary to the replicase, are created in much faster and efficient way, than regular polynucleotides. The more replicases sequences in the solution, the faster new replicases are created, thus the number of molecules available for other reactions is limited.

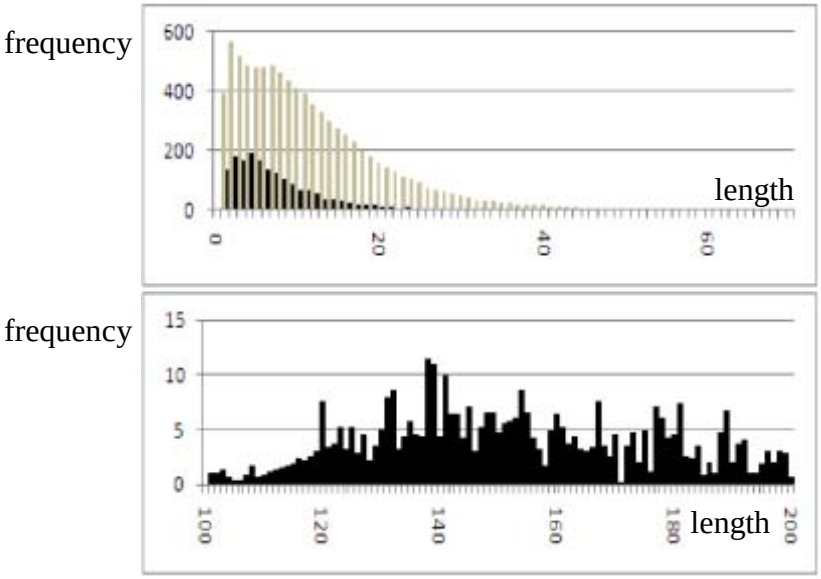


Fig. 6.4:9. RNA molecules lengths without (gray) and with (black) recombination. N = 100,000 and P_{AT} = 0.1
Rys. 6.4:9. Długości molekuł RNA bez (szare) oraz z (czarne) rekombinacją. N = 100,000 i P_{AT} = 0.1

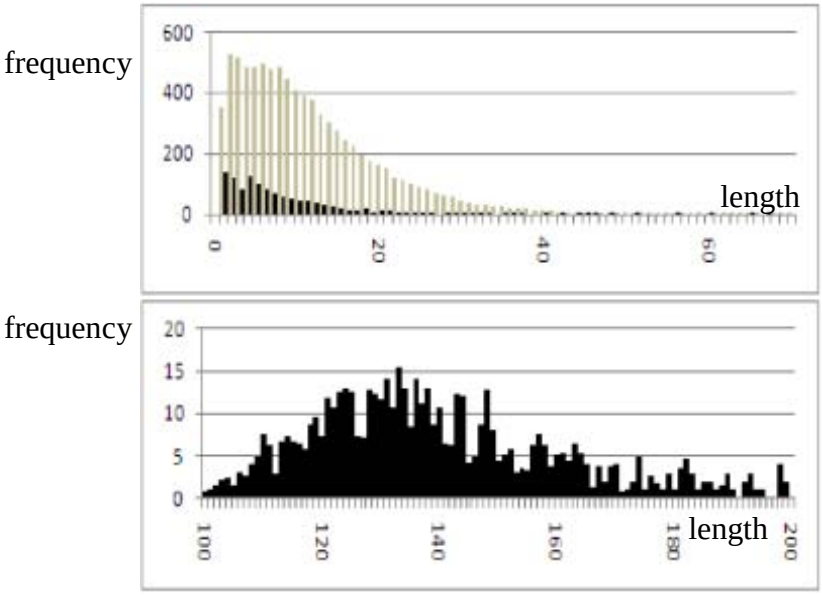


Fig. 6.4:10. RNA molecules lengths without (gray) and with (black) recombination. N = 100,000, P_{AT} = 0.1, and P_{LT} = 0.05

Rys. 6.4:10. Długości molekuł RNA bez (szare) oraz z (czarne) rekombinacją. N = 100,000, P_{AT} = 0.1, and P_{LT} = 0.05

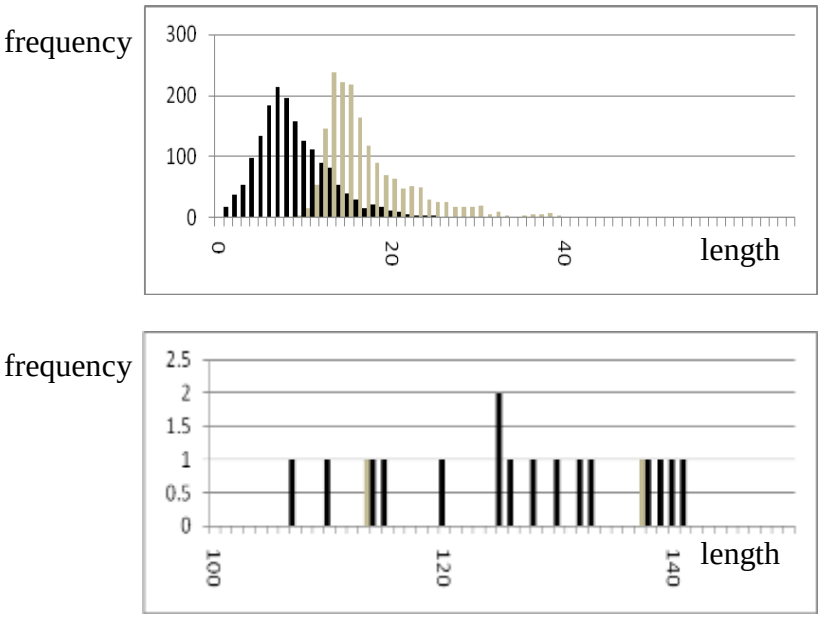


Fig. 6.4:11. RNA molecules lengths in the presence of replicase sequence (9 nt. long) with recombination. Sequences containing replicase sequence (or complementary sequence) are gray, others are Black. N = 50,000

Rys. 6.4:11. Długości molekuł RNA w obecności sekwencji replikazy (o długości 9 nt) z rekombinacją. Sekwencje zawierające sekwencję replikazy (lub sekwencję komplementarną) są oznaczone na szaro, pozostałe na czarno. N = 50,000

6.5. Conclusions

The origin of life is still scientifically open problem, which can be attacked from many different perspectives. One arena, which have given to scientists a better view on this immemorial time is application of computer models of the early life. Section 6.2 of this book addresses the problem of information content threshold in the early stage of RNA-World. This terms refers to the hypothetical stage of the evolution of life which assumes that before emergence of organisms whose genome was based on DNA molecules and enzymatic activities were performed by proteins there existed world of RNA-protospecies in which RNA molecules constituted both the genetic material and enzymes. According to this theory the RNA enzymes, called ribozymes, were required for metabolism and for self-replication.

However, as it was already shown based on *information loss – selection balance* approach, and as it is presented in section 6.2 using branching processes approach, the replication error-rate is a crucial quantity for the maximum information content of the RNA-protospecies. Therefore, one hypothetical ribozyme called RNA replicase is required in the early phase of RNA-World, since it can reduce the mutation rate, and thus, allow for development of genomes with increasing information content. Otherwise, the information would have been lost, and the error catastrophe would have taken place. However, the information preserved in the RNA replicase itself is strongly limited, because in the phase of evolution proceeding the emergence of this ribozyme the replication could not take the advantage of the low mutation rates and yet the evolution of RNA-strands leading finally to the "invention" of replicase had to satisfy the information limiting constraints.

Therefore, RNA replicase would have never been able to evolve if its function could appear only in RNA chains containing large amounts of information. In section 6.2 this problem is considered using model proposed by Demetrius and Kimmel. This model draws the conclusions relaying on the criticality property of branching processes. While utilizing this approach, the author's contribution lies is the introduction into the model the parameters which can be experimentally measured in a test tube. Therefore the estimates of the maximum information content of the primordial RNA-based RNA replicase can be determined using data from biochemical experiments.

Perhaps it is also worth to notice that these estimates corroborate with results obtained with the use of others methods. The method based on a balance between information loss and Darwinian selection predicts equally clearly the difficulties with the origin of self-replicating macromolecules in abiotic environment. Last but not least, the methodology presented in section 6.2 can encourage biochemists for experiments yielding results helpful in the

estimation of the probability of the break of phosphodiester bonds in RNA molecules under conditions feasible on the early Earth.

Up-to-date many models of the early life have been proposed. Some of them rely on mathematical equations (Nowak and Ohtsuki 2008, Ohtsuki and Nowak 2009, Manapat et al. 2009) including those which have only numerical solutions – for example single strand Demetrius/Kimmel model, considered by Cyran (2009b) using the phosphodiester bond break reaction (see section 6.2). Others, formulate conclusions based on computer simulations, like in the case of the compartment model proposed by Niesert et al. (1981) modified by Niesert (1987) and further improved by Myszor and Cyran (2009), as reported in section 6.3.

Extensive simulations of the very first phase of RNA-world have proved that it is feasible to create short RNA strands of the length not exceeding 30 nucleotides (Ma et al. 2007a). It is, however, hard to believe that such short oligonucleotides could catalyze its own replication in selective way. Yet, under the assumptions of non-enzymatic template-directed based RNA-recombination, as shown by Myszor and Cyran (2010) can be even as long as 100 nucleotides or more (see also section 6.4).

The number of nucleotides around 100 is feasible based on the complexity threshold study performed by Cyran (2009b), (see section 6.2) and perhaps is enough for emergence of functional selectivity of the primordial replicase, which is required to amplify the growth of itself and not all unrelated strands. Can be the maximum length of the selective replicase more than 100 nucleotides? In the light of experiments reported in section 6.2, this critical length is dependent on the probability of phosphodiester bond break. What are feasible values of this parameter then?

Certainly r is dependent on the environmental conditions like temperature or the concentration of nucleotides in a solution (the smaller concentration the longer time between replications and thus larger w for equal values of temperature and other environmental parameters such like pH of the solution for example). However, it is possible to obtain experimentally the value of r for given environment. The conditions existing on early Earth, feasible from geological point of view, can be therefore simulated in a test tube and then the model proposed in a paper can be applied with the reliable value of parameter r, as it is already in the case of parameter μ of the order 0.01. When lacking such experiments, the wide range of r from 10^{-5} to 10^{-3} was treated as plausible.

The author expresses his hope, that after performing aforementioned chemical experiments the refinement of the limiting information content can be achieved using proposed here methodology. Until there is a lack of reliable estimates of the phospodiester bond break probability, the discussion with a broad range of its possible values is of some worth. The extremely low value of this probability such as 10^{-5} represents the situation of

substantial concentration of nucleotides and other environmental conditions supporting fast replication.

Whether such conditions are feasible on the early Earth is an open question, but if so, they simplify imagining the evolution of hypothetical primordial RNA replicase to selective replicase catalyzing only its own replication. The complexity threshold for such ribozyme exceeds 500 nucleotides which is probably more than enough to activate the proposed function. However, if r proves to be as large as 10^{-3} or more, then the complexity threshold for selectively working replicase is considerable less and only 170 nucleotides must have been sufficient to constitute such ribozyme.

While this is not impossible, the result would have limited the domain of hypothetical replicases to sequences shorter than 200 nucleotides. Perhaps it would also suggest the double-origin hypothesis in which the replication occurs in biotic condition of metabolizing proto-cells. Such conditions could easier reduce the mutation rate to 10^{-3}, the value that yields complexity threshold well above 10^3 for any considered value of parameter r.

And last but not least, let us mention about the amount of information preserved in the evolving protospecies. In section 6.2, it is shown that before the emergence of the primordial RNA-replicase the amount of information which could be preserved in self-replicating RNA protospecies had to be limited to 10^3 bits. This is twice (because one nucleotide codes for two bits of information) the complexity threshold limit for parameters $\mu = 10^{-2}$, $r = 10^{-5}$. Most probably, i.e. for $\mu = 2\times10^{-2}$ and $r = 10^{-4}$ the amount of information could not exceed 4×10^2 bits. If the RNA-world had ever existed the Nature had to find very information efficient system being able to encode the complex function of emerging primordial RNA-replicase ribozyme having not more than 200 nucleotides i.e. using probably only less than 400 bits of information (Cyran 2009c).

Another simple model of early life proclaims that primitive genes (molecules) were enclosed in compartments (packages) which were submerged in primordial broth. U. Niesert, D. Harnasch and C.Bresch in the article "Origins of Life Between Scylla and Charybdis" explained the basics of the model and predicted that there can be only 3 unlinked types of genes in a package. One of the important factors in the compartment model is the NORM parameter denoting the number of replicated molecules between fission of two packages.

The computer simulations demonstrated that the compartment model with random assortment of genes could not exist without primordial replicase. When this type of gene is not present in a model, parasite mutation rate is too high and the compartment model ends up with a single strand model. The results presented in section 6.3 showed that without NORM variation and in presence of the replicase in a compartment, there might be even 4 different types of genes (100 nucleotides each) for semi-optimistic mutation rates (PMR = 0.001 per

nucleotide, LMR = 0.01, AR = 0.05). However, the NORM variation (whether environmental or individual) leads to limitation in the number of different genes to maximally 3. Higher variation leads to greater reduction in the number of genes. It seems that the environmental variation have greater impact on the maximum number of genes than the within compartment variation on the package level.

It has been also determined (see section 6.4) that the process of non-enzymatic template-directed RNA recombination has an influence on the lengths of strands in the RNA-world. This phenomenon limits the number of shorter sequences, however it might lead to the creation of significantly longer sequences. These longer RNA chains could have catalytic abilities and in longer perspective lead to further RNA strands elongation. What is more, in the presence of RNA recombination process, it was possible to obtain replicase of the same length as in the model without recombination. Emergence of replicase limits the influence of RNA non-enzymatic template directed recombination, at least in the environment with limited supply of constituents.

Based on results presented in section 6.4 it seems that the RNA recombination processes should be seriously considered in explanation of life origins using the RNA-world models. These processes are especially interesting in frozen solutions – in lower temperatures chains are more stable and probability of the oligonucleotide attachment and ligation are rising.

7. GOING BEYOND …

Author's research projects described in the book should be considered in three domains, two of them belonging to information sciences and one to computational biology. Specifically, the book was focused on progress in the areas of AI, computer simulations, and evolutionary genetics. These issues are discussed below.

a) **Advances in AI.** The variety of AI disciplines can be clustered according to two criteria (inspiration, and level of knowledge processing) which, although from different perspectives, define similar groups of methods. The motivation criterion defines a group of biologically inspired methods (ANN, AE), and a group which arose from a formal logic (FS, RS). It is striking that the criterion based on level of knowledge processing yields identical clustering. The low-level processing is attributed to ANNs (connectionist level) and EAs (genetic level), whereas the high-level processing is attributed to FS and RS (rule-based methods). Since we perceive problems at the high-level logic, the rule-based methods do not require the transformation between the human-perceptional and internal-operational levels. Contrary, the connectionist systems need such a transformation and it is really intriguing how well ANNs respond to this challenge. The parallel progress of both these AI groups does not favor in general any of them as significantly more efficient than the other. Therefore, the case studies have to be carried out in order to recognize their appropriateness for particular applications. In the book the comparison between rule-based (CRSA, DRSA, QDRSA) and connectionist methods (PNN, MLP) has been considered in terms of their efficiencies in the problem of the search for natural selection operating at molecular level in genes implicated in human familial cancers. Since this problem is not finished, there is a room for further studies in this context. The classical rough set approach (CRSA) employs indiscernibility relation to generate the granules of indiscernible abstract classes, whereas DRSA relies on dominance relation which generates the granules of dominance cones. The author has proposed in 2009 a novel and original approach called quasi-dominance rough set approach (QDRSA) incorporating concepts of DRSA to CRSA-based granularisation.

Such combination inherits the advantage of ordered attribute domains in DRSA while keeping the advantage of CRSA, the usability of the relative value reducts. There are also limitations which need to be specified to transform QDRSA to a mature approach. The research to be performed by the author in the future is expected to make progress in this respect. Finally, its possible application for screening genes in the search for signatures of natural selection will illustrate the potential of the method.

b) **Progress in computer simulation methods**. Unknown algebraic solutions for time-inhomogeneous BPs as well as solutions for time-homogeneous BPs known solely from the limiting theorems like those proposed by O'Connell, are the reasons for applying computer simulations which can give the insight to the evolution of BP. Traditionally, population geneticists rarely use forward in time simulations of BPs because of high computational (both, time and space) complexity. The reason for lower complexity of algorithms used for simulations backward in time (coalescent methods) is that they process only the lineages observed in a sample. Contrary, algorithms used for time-forward simulations trace the whole genealogy, comprising extinct lineages. Moreover, the simulation of BPs is inherently difficult because of their instability (extinction or growth to enormous size) what constitutes a serious challenge for the effective memory usage. However, these algorithms acquire more attention because of the increase in the computational power and memory sizes of computers, and, somewhat less expected, due to progress in the state-of-the-art within the genome sequencing techniques. It became possible to sequence genes from fossils of the extinct species (Neanderthal genome project is the sound example of the enormous progress in the sequencing technology). Since such genetic data require observing their decay leading ultimately to the elimination from the gene pool (the effect of genetic drift is algebraically tractable in Wright-Fisher model only for simple genealogies), the forward in time simulations open new perspectives. Problems concerning genetic drift of Neanderthal mtDNA loci, have been presented in the book. Similarly, the maximum number of different genes in the RNA-world early compartments was studied based on development and implementation of an advanced simulation-based model. The novelty of this simulation model lies in studying complex stochastic effects associated with the interplay of environmental (generation-to-generation) and individual (cell-to-cell) variations of NORM parameter during simulated evolution. There is also potential for the introducing BP criticality criterion which will be exploited in further studies.

c) **Advances in evolutionary genetics**. These include the development of author's original multi-null-hypotheses method and application of AI technologies in the search for natural selection. MNH is a novel technique whose concept requires further studies: only a fraction of possible applications has been described. The critical values of neutrality

statistics tested against modified nulls in MNH have to be estimated separately for each gene and/or population considered. Therefore, despite the potential for high accuracy the application of MNH as a screening technique is doubtful. This is where progress in AI can help. The accurate MNH results obtained from neutrality tests, such as Tajima's *T*, Fu and Li's *D** and *F**, Kelly's Z_{nS}, Wall's *Q* and *B*, and Strobeck's *S* can be used as an expert knowledge. The generalization of this knowledge is expected to be a basis for screening procedure, the more, that there is a potential for inventing the automatic selection tester. Additionally, the book described studies concerning the field of *H. sapiens* evolution by presenting methods for estimating possible admixture levels of Neanderthal mtDNA. These estimates are relevant for the progress of the state-of-the-art in the field, as complementary to estimates expected to be obtained based on Neanderthal nuclear DNA (being sequenced in the Neanderthal genome project). Advances in methods estimating information amount in RNA-world protospecies demonstrate the potential of information theory in the domain of the origins of life. The models studied are based on parameters which can be estimated using biochemical experiments of RNA evolution in a test tube (mutation rate) and there is also a potential for experimental ascertainment of the second parameter (probability of phosphodiester bond break). Inventions in that matter demonstrate the growing interest in information sciences within the field concerning complexity threshold in the early life.

The two regions of biological evolution – origins of life and origins of humans – are situated among the most fundamental issues influencing scientific understanding of Nature. Currently, for the first time in the history of science, these fundamental problems are tried to be solved based on a huge amount of empirical, genomic data, and the enormous number of biochemistry experiments demonstrating in a test tube relevant phenomena operating at molecular level. These features, characterizing the state-of-the-art in evolutionary genetics are the reasons for research focused on the use of intensive computer simulations and AI methods, supporting researchers with powerful modeling tools and facilitating understanding of genetic data. Short analysis which follows, explains it further.

The completion of the Human genome project is a symbolic caesura starting the post-genomic era, characterized by huge amount of genetic data and a permanent need for data processing and understanding techniques which go beyond the classical understanding of bioinformatics. This situation has been strengthen after Common Chimpanzee genome project was initiated. In 2004, a preliminary analysis of 7600 genes shared between the two genomes confirmed that genes such as the forkhead-box P2 transcription factor, involved in speech development, are different in the human and chimpanzee lineages. Several other genes involved in hearing were also found to have changed during human evolution, suggesting that

natural selection operating at molecular level has shaped human language-related behavior. Differences between humans and chimpanzees estimated to be 10 times the typical difference between pairs of humans constitute the basis for tuning molecular clock used in studying human evolution.

The Neanderthal genome project launched in 2006 is expected to yield roughly 3.2 billion base pairs of the Neanderthal genome. From the very beginning it is a joint European and US research. The project was launched in July 2006 by the Max Planck Institute for Evolutionary Anthropology in Germany and many research institutes in the United States announced that they would be sequencing together the Neanderthal genome over the next several years. The most prominent genetic centers in the United States are interested in cooperation with institutions of the European Research Area because ancient DNA from Neanderthals fossils was found solely in Europe. Among other, the researchers extracted the DNA from the femur bone of a 38,000-year-old male Neanderthal specimen from Vindija Cave, Croatia, and also other bones were found in Spain, Russia, and in Germany. According to preliminary results modern human and Neanderthal DNA appear to be 99.5% identical (compared to humans sharing around 95% of their genes with chimpanzees). The conclusions of two research teams studying the same Neanderthal sample, published by Richard Green's team in *Nature* (Pennisi 2006, 2007), and Pääbo and Roobin's group in *Science* (Noonan *et al.* 2006), were received with some criticism, mainly surrounding the issue of the admixture of Neanderthals to the genome of *H. sapiens*. The possibility of admixture is strengthen by the fact that the speech-related gene FOXP2 with the same mutations as in modern humans was discovered in ancient DNA in the El Sidron 1253 and 1351c specimens. It suggests that Neanderthals might have shared some basic language capabilities with *H. sapiens* what could support genetic exchange to the extent estimated preliminarily in 2005 by the author (Cyran and Kimmel 2005) based on mtDNA record. In February 2009, the Planck Institute's team, led by geneticist Svante Pääbo, announced that they had about 63% of the entire base pairs. An early analysis of the data suggested "no significant trace of Neanderthal genes in modern humans". In this context, simulating in the future the effect of genetic drift using BP model expected to be more accurate than that of 2005, is correlated with the research based on nuclear DNA being sequenced in Neanderthal genome project.

Current theories concerning the origin of life fall into two groups defined by Dyson (1999) in his famous book The o*rigins of Life*. The first group assumes that the transition form abiotic to biotic world occurred with the emergence of self-replicating RNA molecules and is referred to as RNA-world hypothesis. This most commonly accepted hypothesis requires the existence of the RNA-replicase ribozyme the search of which is described by McGinness and Joyce (2003). The evolution of new genes after appearance of the RNA-replicase is challenged by instability of Eigen's hypercycles composed of many genes

supporting cyclically their replication. An alternative approach, proposed by Niesert *et al.* (1981) as a compartment model with random segregation of genes, proved to be stable for very limited number of genes. Significant advance in the RNA-world theory has been done by Ma et al. (2007a) who performed intensive computer simulations demonstrating the emergence of the auto-catalytic and self-replicating activity of RNA oligonucleotides. Another relevant computer simulation-based study was reported by Baaske *et al.* (2007) who observed the extreme accumulation of nucleotides in simulated hydrothermal pores. The second group of hypotheses derives life from the biochemistry of amino acids and their polymers, proteins. This group encompasses such theories like Dyson's theory of double origin which requires at least 8-10 types of monomers for emergence of the first auto-catalyzing protocells and therefore excludes from this role nucleotides, or theories described by Rode *et al.* (2007) assuming that salt-induced peptide formation (SIPF) reaction could have been the crucial step from chemistry towards biology.

In this context it should be noticed that the problem of complexity threshold, considered in section 6.2 is equally important for both groups of theories although for each of them the acceptable value of complexity threshold is different Therefore the reliable estimate of this threshold based on methodology proposed in the book could favor one or the other group, or at least predict the limits for the length of newly arisen genomes and in that matter contribute to revealing the mystery of life. Note, that the studies described in section 6.4 are not over yet. In the next step there is plan to incorporate into the model also other types of RNA recombination and investigate influence of these processes on model's ability to create long RNA molecules, and polynucleotides containing replicase sequence.

Finally, let us focus on some general issues concerning the information sciences, and in particular, the artificial intelligence, as they might appear in the future. The first problem is the fundamental mode of operation of information processing systems. The second problem is that of complexity, however, considered here from the opposite side, as compared to the study in section 6.2. While section 6.2 considered maximum complexity of protoorganisms which not necessarily had to be degraded by the error catastrophe, here, after von Neumann, it will be discussed the minimum complexity required for evolution understood as a production of more and more complex individuals. The reader will be left with implications of the fact that at the same time the complexity of any self-replicating system, must satisfy both these bounds. The third problem tackles the strong artificial intelligence, and hence, in some parts, it goes beyond the science and addresses philosophical views on what the intelligence is, and can it be present in artificial automata of the future. These important issues, signaled in this paragraph, are briefly discussed in what follows.

The operation of any information processing system can be classically described in terms of logic statements, such as conjunctions, disjunctions, or negations. Such description,

however, even if correct from logical point of view, cannot be treated as a complete model of operation of the real system because real systems are composed of elements with low but non-zero probabilities of the malfunction. The probability of erroneous operation of the whole system can be reduced by increasing the information redundancy in it, yet still it will operate with arbitrarily low, but non-zero probability of the error. In the context we are primarily focused on in this book, it is probably worth to say that such erroneous processing of the genetic information in living organisms (called mutation) allowed for the evolution from at least the time when self-replication of the precursors of modern genes occurred. Whether this event was equivalent to the origin of life is a problem discussed in Section 6.1. Here, we are more concerned with such theory of information processing which can adequately describe also the operation of real (erroneous) automata.

The necessity of such theory was postulated 60 years ago by von Neumann (1951), who claimed less combinatorial and more analytical nature of it. After 60 years separating us from his lecture and despite some exceptions, the strong bias towards combinatorial and not analytical treatment of the information processing systems seems to be omnipresent also in modern informatics. These rare exceptions include that part of theoretical physics which comes relatively close to notions present in manipulating and measuring the information, such as thermodynamics inherited from Boltzmann and further developing also in a context of information processing. The notion of informational entropy is the most prominent, but this theory serves also as a model in non-classical information processing systems like neural networks discussed in section 2.2.1. Another theory which can be viewed as more analytical theory of information is the informational macrodynamics (Lerner 2003), and to some extent theory of DNA computing (Paun et al. 1998).

When von Neumann (1951) was saying about greater complexity of natural organisms as compared to artificial automata he considered *prima facie* modern life. However, the concept of complication treated by him not only quantitatively, but after achieving some level, also qualitatively, indicates that all organisms (not only modern) must posses at least that level of complexity to self-reproduce and evolve. The same can be said about artificial self-reproducing automata and about natural living organisms. Von Neumann was well aware of the many important differences between the two, but in what concerns processing information and problem of self-reproduction, he clearly associated complexity with organization, and concluded, that "*complication on its lower levels is probably degenerative, that is, that every automaton can produce other automata will only be able to produce less complicated ones. There is, however, a certain minimum level where this degenerative characteristic ceases to be universal. At this point automata which can reproduce themselves, or even construct higher entities, become possible. This fact, that complication, as well as organization, below certain minimum level is degenerative, and beyond that level can become*

self-supporting and even increasing, will clearly play an important role in the future theory of the subject". Although stated by not a chemist, neither by biologist, this conclusion should be more seriously taken into account in theories trying to explain the origins of life.

On the other end of the evolution, there is located an emergence of modern humans and the raise of human intelligence. Is this intelligence something special? To answer, Crevier (1993) cites Minsky's words "*if the nervous system obeys the laws of physics and chemistry, which we have every reason to suppose it does, then .... we ... ought to be able to reproduce the behavior of the nervous system with some physical device".* However, this argument has much longer history. In fact, it was first introduced by McCulloch and Pitts (1943) and later it was given by, among others, Moravec (1988). Kurzweil (2005) is convinced that a complete brain simulation using computers will be possible in 2029. While, giving such exact date in predicting future, seems to be highly irrational, the importance of Kurzweil's estimate lies in the time-proximity to the predicted event. Indeed, the next decades, rather than centuries, should give us the answer to one of the most fundamental issues concerning human mind – is it a unique product of human specific evolution, or, as suggested by supporters of strong AI, it is simply a product of the large enough complexity.

Notably, some experiments on a large scale have already been performed. Izhikevich and Edelman (2008) report an interesting example of modeling, on a cluster of 27 processors, a thalamocortical system comparable with a size of the human brain with approximately 10^{11} neurons. This experiment was executed three years earlier, in 2005, however, it should be stressed that it was a non-real-time simulation, which required 50 days in order to model dynamics of the brain activity lasting only 1 second. Currently, due to such and similar experiments, the majority's view is that brain simulations are theoretically possible.

This opinion is supported even by Dreyfus (1972) who is known to criticize the artificial intelligence understood as generating computer programs that can embody consciousness. Similarly, Searle (1980) disagreeing with hopes in the success of such approach, writes: "*What we wanted to know is what distinguishes the mind from thermostats and livers*". Therefore, for Searle (1999), the difference between weak artificial intelligence and strong artificial intelligence is as fundamental as between "liver and thermostat" as opposed to "mind". This difference is consciousness, which for such researchers as I. Aleksander (see for example Aleksander 2008), S. Franklin (see Franklin 1997), R. Sun (see Sun 2002), and P. Haikonen, is the necessary component of intelligence. In this context, Heikonen (2003) writes: "*the brain is definitely not a computer. Thinking is not an execution of programmed strings of commands. The brain is not a numerical calculator either. We do not think by numbers*".

Russell and Norvig (2003) in their review write, that the most computer scientists take the weak AI hypothesis for granted, and they are not much interested in studies trying to prove or

disprove the strong AI hypothesis. While this is maybe true, there are some, including the author, who are interested in issues raised by the strong AI hypothesis, as belonging to fundamental problems of Nature. Perhaps these problems can be addressed by interdisciplinary approaches of modern physics and neuroscience. Some well known neurosurgeons (K.H. Pribram), or physicists (R. Penrose) are arguing that quantum theory can provide the foundations for explaining the consciousness.

Pribram (1991) is the author of holonomic brain theory, which is inspired by holography. Hameroff and Penrose (1996) using self-organized objective reduction phenomenon, developed the Orch OR theory of quantum consciousness. Note, that although this theory supports view that classical physics is intrinsically incapable of explaining consciousness, none of the quantum mechanical theories has been experimentally confirmed. Therefore, many scientists and philosophers are unconvinced as to the essential role of the quantum phenomena in creation of the consciousness.

Let us now consider one of the most famous arguments against strong AI hypothesis that machines can consciously think. This thought experiment, known as the "Chinese Room" was proposed by Searle (1980) with a goal to prove that machines cannot be conscious even if they pass the Turing test. Assuming that some computer program has passed the test proposed by Turing (1950), and moreover, that it can speak Chinese, let the instructions of this program be written on a paper cards and given to a man, who does not understand Chinese. Let this man be closed in a room with a slot for exchanging messages with a person being outside the room, who fluently speaks Chinese. From the outside world it seems that the system composed of a room and a man processing instructions of the program typed on the cards can speak Chinese. However, Searle (1980) argued that nobody (or nothing) in a system understands the meaning of what has been spoken. Therefore, he concluded, Chinese Room (or any other symbolic AI system, which passed Turing test) cannot be aware of the sense of what has been said. The consciousness and mental states are reserved only for mind, which requires not only complexity, but also physical and chemical properties characteristic for human brain.

The responses to the Chinese Room argument are given from many different perspectives (Cole 2004). For example, the so called "virtual mind reply" (a) and the "systems reply" (b) point out that the system, including the man, the program, the room, and the cards, does understand Chinese. Another response argues that the man in the room would probably require millions of years and extremely huge number of auxiliary cards to respond to a simple Chinese question. Such arguments, are together referred to as "speed, power and complexity reply" (c). The "robot reply" (d) argues that the Chinese Room needs eyes and hands to understand truly, and the "brain simulator reply" (e) focuses on situation when the program typed in cards simulates the nerve signals of an actual Chinese speaker, so the man in the

room would be simulating an actual brain. Next, the so called “other minds reply” (f) points out that since it is not easy to prove that people are "really" thinking (compare also with the Turing’s *polite convention*), so it is also hard to decide in the case of machine. And finally, the “epiphenomena reply” (g) shows difficulties in the Searle’s belief that natural selection created neurons, whose “casual properties” responsible for emergence of consciousness, are epiphenomenal, i.e., they make no difference to behavior (according to Searle, both conscious humans and unconscious machines can pass the Turing test, i.e. their behavior would be identical).

However, the replies to the Chinese Room argument are also arguable.

a) Virtual mind reply relies on observation that computers can be organized in many layers of virtual machines and one physical machine (hardware), and each of this machine performs completely different information processing. Therefore, the Chinese Room as a whole can have another level of consciousness, which understands Chinese (additionally to consciousness of the man inside a room). However, the fact that the symbolic processing can be implemented in virtual machines does not imply that the same is true with the consciousness. In fact, there is not a single experiment which would prove that the consciousness can appear in virtual machines.

b) Systems reply claims, that since systems can acquire some new quality (which is not a simple sum of its components), therefore the Chinese Room as a whole can understand what has been spoken. However, it has not been demonstrated that consciousness can appear in such a way. In fact, if the consciousness of the Chinese Room has arisen, there would be, at the same time, the confirmation of the “virtual minds reply” since it would be the second level of consciousness built-up over consciousness of a man in the room. Yet, nothing similar was confirmed to emerge in similar situations. All experiments confirm the opposite, that the only intelligence in Chinese room is that of the man.

c) Speed, power, and complexity reply states that consciousness can occur only in conditions of high enough speed/power/complexity. Since the man in the room is a very slow “processor” he cannot understands Chinese when implementing the algorithm. If he was fast enough he would understand. However, this latter statement rather difficult to prove/disprove.

d) Root reply is non convincing, as we have examples of humans who are visually or physically impaired and still conscious. Therefore, it is hard to imagine why “eyes and hands” would be crucial for disprove the Chinese Room argument.

e) Brain simulator reply is a special case of virtual mind reply. It was posed to prove that if the algorithm was written in terms of nerve signals simulators then the processor

implementing this algorithm should have consciousness. However, such argument assumes (without any experimental evidence) that for consciousness does not require specific hardware (i.e. brain) and can emerge in virtual system. The reservations are identical to those mentioned in (a).

f) Other minds reply tries to extend the Turing's polite convention (if we cannot be 100% sure that others think, we should assume that they think) to any systems. While it is really a hard problem how to define a test which will be able to verify whether some system/machine is conscious (Turing's test will not be enough) it is equally difficult to accept the view the Chinese Room has consciousness (different that that of a man) only because we cannot directly verify the opposite. Some positive argument would be required.

g) Epiphenomena reply says that it is evolutionarily incredible to maintain that consciousness would arise epiphenomenally (and therefore it was selectively neutral), and such conclusion can be derived from hypotheses stating that the behavior of unconscious entities and conscious individuals cannot be discern based on any external test (such as Turing test). However, if consciousness in biological life is indeed evolutionarily favorable (what it is hard not to believe in) it does not imply that behavior of some unconscious machine and a conscious man, necessarily must be distinguishable. The point is that machines do not arise by evolution, and they are specifically designed to mimic operation of a conscious and intelligent man. Most probably, they would require much more computational power than that of the human brain to achieve this goal, however they will be constructed using finite, but not limited by evolutionary cost, resources. And hence, they would never emerge by evolution, because of not efficient management of their potential computational abilities. In other words, assuming the same level of computational power, the conscious individual would have been selectively preferred over unconscious, what does not deny, that unconscious creature (machine) could have similar fitness as conscious if supplied (by a constructor) with much greater computational abilities.

After presenting the Chinese Room argument with critical replies and author's comments on these replies, let us focus on the next philosophical issue, the emotions and self-awareness of machines. After loosing the chess match with Deep Blue 2 computer, the Chess World Master G. Kasparov nervously mentioned that even if the computer won, it had completely no satisfaction. Not only the satisfaction was absent in Deep Blue 2, but also the awareness of the victory, and of course the awareness of self-existence. Will it change in the future? Will be the (next)…(next) generation computers self-aware?

Turing (1950) wrote "*I do not wish to give the impression that I think there is no mystery about consciousness ... but I do not think these mysteries necessarily need to be solved before we can answer the question*" of whether machines can think. Methodological reductionism, involved in the above statement, led Turing to reduce the problem of machine's self-awareness to such answers, to questions "*can a machine be the subject of its own thought*?" and "*can it think about itself*?", which did not tackle a problem of consciousness. In the light of this methodological reductionism it is easier to understand that Turing (1950) indicated a computer with running debugger (i.e. a program which can report on its own internal states), as an example of self-aware machine.

Note, that in the light of this reductionism "reporting on its own states" is equivalent to "thinking about itself", since "reporting" and "thinking", can be treated as synonyms when considered without taking into account consciousness. However, it is worthwhile to mention, that under such methodologically assumed equivalence, "thinking" of the debugger about the host computer is an activity of exactly the same nature as "thinking" of the database management software about the facts stored in the database. Turing (1950), did not want to give the definition of thinking, however, it is clear from his paper (not only from the cited above sentence) that consciousness was not considered by him as a *condition sine qua non* of thinking.

Note, that Turing's methodological reductionism should be always considered in a context of his words the "*mystery about consciousness*". When this mystery is contemplated, then the possibility of creation of self-aware machines is still an open question (despite existence of debuggers), to the same extent, as it is open problem whether a database management system will ever be able *to understand* the data (despite it can report them perfectly in currently available systems). Interestingly, Turing (1950) wrote also that in 2000 there should be computers with memories of the order of gigabytes and that such machines will be able to successfully pretend humans in abilities of natural language processing to such extent, that during typical, five-minute-long conversation, the computer would become unrecognized in around 70% cases. While his first intuition (that about technical progress) proved to be underestimated, the second one (that about naturally speaking computers) is an overestimate even in 2010 when we have at our disposal memories with terabytes and more.

So, 60 years after Turing's seminal paper the strong AI hypotheses is neither proved or disproved. Kurzweil (2005) speculates that machines implementing strong AI paradigm will be available in 20 years from now, but Searle (1980) is principally pessimistic about possibility of creation strong AI systems, at least implemented as symbol processing machines. Who is right? Each of us has the personal guess. Up to now, it has to be intuition and guess, because neither Chinese Room argument, nor the replies to it, are solid scientific

proofs, which would necessitate their acceptance. The only proof of the strong AI claims will be a hypothetical meeting with a conscious machine.

But, will it be recognized as such?

It will perhaps pass Turing's game…

How will we then address what Turing has called

the mystery about consciousness?

We have to go beyond now, to be prepared for future…

We would not be human beings, if we were not going beyond…would we?

BIBLIOGRAPHY

1. Adachi J., Hasegawa M. (1995): Improved dating of the human-chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. 40, p. 622÷628.
2. Adams M. D., McVey M, Sekelsky J. J. (2003): Drosophila BLM in double-strand break repair by synthesis-dependent strand annealing. Science 299, p. 265÷267.
3. Agrafioti I., Stumpf M. P. H. (2007): SNPSTR: a database of compound microsatellite-SNP markers. Nucleic Acids Res. 35 (supplement 1), p. D71÷D75.
4. Akashi H. (1995): Inferring weak selection from pattern of polymorphism and divergence at 'silent' sites in Drosophila DNA. Genetics 139, p. 1067÷1076.
5. Aleksander I. (2008): Machine consciousness. Scholarpedia 3(2): p. 4162÷4162.
6. Angeline P. J. (1997a): Evolutionary Computation Models – Representations – Parse trees. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C1.6:1÷C1.6:3.
7. Angeline P. J. (1997b): Evolutionary Computation Models – Search operators - Mutation – Parse trees. In In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C3.2:9÷C3.2:10.
8. Angeline P. J., Fogel D. B. (1997): Evolutionary Computation Models – Representations – Other representations. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C1.6:1÷C1.6:3.
9. Azuaje F. (2003): Genomic data sampling and its effect on classification performance assessment. BMC Bioinformatics 4(1), p. 5-16.
10. Baaske Ph., Weinert F. M., Duhr S., Lemke. K. H., Russel M. J., Braun D. (2007): Extreme accumulation of nucleotides in simulated hydrothermal pore systems. Proc. Natl. Acad. Sci. USA 104, p. 9346÷9351.
11. Bachtrog D., Charlesworth B. (2001): Towards a complete sequence of the human Y chromosome. Genome Biol. 2(5), reviews 1016.1÷ reviews 1016.5.

12. Bamshad M. J., Mummidi S., Gonzalez E., Ahuja S. S., Dunn D. M., et al. (2002): A strong signature of balancing selection in the 5' cis-regulatory region of CCR5. Proc. Nat. Acad. Sci. USA 99(16), p. 10539÷10544.
13. Bäck T. (1997a): Evolutionary Algorithms and Their Standard Instances - Introduction. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. B1.1:1÷B1.1:4.
14. Bäck T. (1997b): Evolutionary Computation Models – Representations – Binary strings. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C1.2:1÷C1.2:3.
15. Beamish H., Kedar P., Kaneko H., Chen P., Fukao T., et al. (2002): Functional link between BLM defective in Bloom's syndrome and the ataxia-telangiectasia-mutated protein, ATM. J. Biol. Chem. 277, p. 30515÷30523.
16. Berfanger D. M., George N. (1999): All-digital ring-wedge detector applied to fingerprint recognition. App Opt . 38 (2), p. 357÷369.
17. Berfanger D. M., George N. (2000): All-digital ring wedge detector applied to image quality assessment. App Opt. 39(23), p. 4080÷4097.
18. Birgmeier M. (1996): Evolutionary programming for the optimization trellis-coded modulation schemes. Proceedings of Fifth Annual Conference on Evolutionary Programming, San Diego, CA, Fogel L. J., Angeline P. J., Bäck T. (eds), Cambridge, MA, MIT Press.
19. Bjorklund M. (2003): Test for a population expansion after a drastic reduction in population size using DNA sequence data. Heredity 91(5), p. 481÷486.
20. Bobrowski A., Kimmel M. (2004): Asymptotic behavior of joint distributions of characteristics of a pair of randomly chosen individuals in discrete-time Fisher-Wright models with mutations and drift. Theoretical Population Biology 66, p. 355÷367.
21. Bolc L., Dziewicki K., Rychlik P., Szałas A. (1995): Wnioskowanie w logikach nieklasycznych – podstawy teoretyczne. Problemy Współczesnej Nauki – Teoria i Zastosowania, Informatyka, Akademicka Oficyna Wydawnicza PLJ, Warszawa.
22. Bonnen P. E., Story M. D., Ashorn C. L., Buchholz T. A., Weil M. M., Nelson D. L. (2000): Haplotypes at ATM identify coding-sequence variation and indicate a region of extensive linkage disequilibrium. Am. J. Hum. Genet. 67, p. 1437÷1451.
23. Bonnen P. E., Wang P. J., Kimmel M., Chakraborty R., Nelson D. L. (2002): Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res. 12, p. 1846÷1853.

24. Booker L. B. (1997): Evolutionary Computation Models – Search Operators – Recombination – Binary Springs. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C3.3:1÷C3.3:10.
25. Bourbakis N. G. (2002): Emulating human visual perception for measuring difference in images using an SPN graph approach. IEEE Transactions on Systems, Man, and Cybernetics, Part B 32(2), p. 191÷201.
26. Briggs A. W., Good J. M., Green R. E., Krause J., Maricic T., Stenzel U., Lalueza-Fox C., Rudan P., Brajkovi D., Kuan E., Gui I., Schmitz R., Doronichev V. B., Golovanova L. V., de la Rasilla M., Fortea J., Rosas A., Pääbo S. (2009): Targeted retrieval and analysis of five Neandertal mtDNA genomes. Science 325, p. 318÷321.
27. Brown P., Sutikna T., Morwood M. J., Soejono R. P., Jatmiko, Wayhu Saptomo E., Rokus Awe Due (2004): A new small-bodied hominin from the late Pleistocene of Flores, Indonesia. Nature 431, p. 1055÷1061.
28. Budowle B., Chakraborty R. (2001): Population variation at the CODIS core short tandem repeat loci in Europeans. Legal Medicine 3, p. 29÷33.
29. Budowle B., Shea B., Niezgoda S., Chakraborty R. (2001): CODIS STR Loci Data from 41 Sample Populations. Journal of Forensic Sciences 5, p. 453÷489.
30. Cajavec B. (2002): Getting started with molecular biology. In Proc. of the School of Population Dynamics, Będlewo, Poland, 13÷30.
31. Casasent D, Song J. (1985): A computer generated hologram for diffraction-pattern sampling. Proc SPIE 523, p. 227÷236.
32. Cavalli-Sforza L. L., Bodmer W. F. (1971): The Genetics of Human Populations, Freeman, San Francisco.
33. Cebrat S., Pekalski A. (2004): The Role of Dominant Mutations in the Population Expansion. International Conference on Computational Science, p. 765÷770.
34. Ciemniewski Z., Letkiewicz S., Cyran K. (1997): Connectionist approach in diagnosis support systems on the basis of feed-forward ANN giving prognosis in urology and cardiology. Proc. International Workshop: Biomedical Engineering and Medical Informatics, Gliwice, Poland, p. 100÷104.
35. Chakraborty R. (1986): Gene Admixture in Human Populations: Models and Predictions. Yearbook of Physical Anthropology 29, p. 1÷43.
36. Cochrane J. C., Strobel S. A. (2008): Riboswitch effectors as protein enzyme cofactors. RNA 14, p. 993÷1002.
37. Cohen M. A., Grossberg S. (1983): Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. IEEE Transactions on Systems, Man, and Cybernetics 13, p. 815÷826.

38. Cole D. (2004): The Chinese Room Argument. In Zalta, Edward N., The Stanford Encyclopedia of Philosophy.
39. Cortez D., Wang Y., Qin J., Elledge S. J. (1999): Requirement of ATM-dependent phosphorylation of Brca1 in the DNA damage response to double-strand breaks. Science 286, p. 1162÷1166.
40. Crevier D. (1993): AI: The Tumultuous Search for Artificial Intelligence, NY: Basic Books, New York.
41. Cyran K., Letkiewicz S., Wojciechowski P., Kołoczek D. (1997): Use of neural network to recovery prognosis for patients with renal cancer [in Polish]. ZN Pol. Śl., Informatyka 33, p. 185÷202.
42. Cyran K., Podeszwa T. (1999): Wykorzystanie HMM oraz NN do rozpoznawania kontekstowego w przetwarzaniu mowy, ZN Pol. Śl., Informatyka 37, p. 7÷23.
43. Cyran K. A., Jaroszewicz L. R. (2000): Rough set based classification of interferometric images. In Jacqout P., Fournier J. M. (eds.) Interferometry in Speckle Light: Theory and Applications, Berlin Heidelberg New York, Springer, p. 413÷420.
44. Cyran K. A., Mrózek A. (2001): Rough sets in hybrid methods for pattern recognition. Int. J. Intell. Syst. 16 (2), p. 149÷168.
45. Cyran K. A., Jaroszewicz L. R., Niedziela T., Merta I. (2001a): Concurrent signal processing in optimized hybrid CGH-ANN systems. Opt. Appl. 31(4), p. 675÷689.
46. Cyran K. A., Jaroszewicz L. R., Niedziela T. (2001b): Neural network based automatic diffraction pattern recognition. Opto-electronics Rev. 9 (3), p. 301÷307.
47. Cyran K. A, Niedziela T., Jaroszewicz L. R. (2001c): Grating-based DOVDs in high-speed semantic pattern recognition. Holography 12(2), p. 10÷12.
48. Cyran K.A., Stańczyk U., Jaroszewicz L. R. (2002): Subsurface stress monitoring system based on holographic ring-wedge detector and neural network. In McNulty G. J. (ed.) Quality, Reliability and Maintenance, Bury St Edmunts London, Professional Engineering Publishing, p. 65÷68.
49. Cyran K. A (2003): PLD-based rough classifier of Fraunhofer diffraction pattern. Proc. Int. Conf. Comp. Comm. Contr. Tech., Orlando, FL, USA, p. 163÷168.
50. Cyran K. A., Kimmel M. (2004a): Distribution of time to coalescence under stochastic population growths: application to MRCA dating. In Gramada A., Bourne Ph. E. (eds.) Currents in Computational Molecular Biology 2004: RECOMB 2004, San Diego, USA, p. 11÷12.
51. Cyran K. A., Kimmel M. (2004b): Robustness of the dating of the most recent common female ancestor of modern humans. Proc. 10th National Conference on Application of Mathematics in Biology and Medicine, Święty Krzyż, p. 19÷24.

52. Cyran K. A., Polańska J., Kimmel M. (2004): Testing for signatures of natural selection at molecular genes level. Journal of Medical Informatics and Technologies 8, p. 31÷39.
53. Cyran K. A. (2005a): Combining rule based and connectionist approaches in a diffraction pattern recognition. Proc. Artificial Intelligence Studies 2(25), p. 149÷157.
54. Cyran K. A. (2005b): Integration of classifiers working in discrete and real valued feature space applied in two-way opto-electronic image recognition system. Proc. of the fifth IASTED International Conference on Visualization, Imaging, and Image Processing, Benidorm, Spain, p. 592÷597.
55. Cyran K. A., Kimmel M. (2005): Interactions of Neanderthals and modern humans: what can be inferred from mitochondrial DNA. Mathematical Biosciences and Engineering 2(3), p. 487÷498.
56. Cyran K. (2007a): Rough sets in the interpretation of statistical tests outcomes for genes under hypothetical balancing selection. Lecture Notes in Artificial Intelligence 4585, p. 716÷725.
57. Cyran K. A. (2007b): Mitochondrial Eve dating based on computer simulations of coalescence distributions for stochastic vs. deterministic population models. Proc. 7th WSEAS International Conference on Systems Theory and Scientific Computations, Athens, Greece, p. 107÷112.
58. Cyran K. A. (2007c): Comparison of neural network and rule-based classifiers used as selection determinants in evolution of feature space. WSEAS Trans. on Systems 6(3), p. 549÷555.
59. Cyran K. A. (2007d): Simulating branching processes in the problem of Mitochondrial Eve dating based on coalescent distributions. International Journal of Mathematics and Computers in Simulation 1(3), p. 268÷274.
60. Cyran K., Stańczyk U. (2007a): Indiscernibility relation for continuous attributes: application in image recognition. Lecture Notes in Artificial Intelligence 4585, p. 726÷735.
61. Cyran K. A., Stańczyk U. (2007b): Stochastic simulations of branching processes: Study on complexity threshold of RNA-world species. Proc. XXXVI Ogólnopolska Konferencja Zastosowań Matematyki, Zakopane, Poland, p. 19÷22.
62. Cyran K. A. (2008a): Complexity threshold in RNA-world: computational modeling of criticality in Galton-Watson process. Proc. 8th WSEAS International Conference on Applied Computer Science, Venice, Italy, p. 290÷295.
63. Cyran K. A. (2008b): Modified indiscernibility relation in the theory of rough sets with real-valued attributes: application to recognition of Fraunhofer diffraction patterns. Transactions on Rough Sets IX, Lecture Notes in Computer Science 5390, p. 14÷34.

64. Cyran K. A., Myszor D. (2008a): Coalescent vs. time-forward simulations in the problem of the detection of past population expansion. International Journal of Applied Mathematics and Informatics 2(1), p. 10÷17.
65. Cyran K. A, Myszor D. (2008b): Neural networks and statistical tests for detection of population expansion. Proc. 2nd European Computing Conference, Malta, p. 222÷227.
66. Cyran K. A., Myszor D. (2008c): New artificial neural network based test for the detection of past population expansion using microsatellite loci. International Journal of Applied Mathematics and Informatics 2(1), p. 1÷9.
67. Cyran K. A. (2009a): PNN for Molecular Level Selection Detection. Lecture Notes in Electrical Engineering 27, p. 35÷41.
68. Cyran K. A. (2009b): Information amount threshold in self-replicating RNA-protospecies: branching processes approach. International Journal of Mathematics and Computers in Simulations 3(1), p. 20÷29.
69. Cyran K. A. (2009c): Problem ilości informacji w protoorganizmach świata RNA. Proc. Konferencja Chrzescijańskiego Forum Pracowników Nauki: Nauka-Etyka-Wiara, Jastrzębia Góra, Poland, p. 34÷49.
70. Cyran K. A. (2009d): Quasi Dominance Rough Set Approach in Testing for Traces of Natural Selection at Molecular Level. In Cyran K. A. et al. (eds.) Advances in Intelligent and Soft Computing, Springer, 59, p. 163÷172.
71. Cyran K. A., Niedziela T. (2009): Optoelectronic method of pattern recognition of motor vehicles in spatial frequency domain. Archives of Transport 21(1-2), p. 27÷47.
72. Cyran K. A. (2010): Classical and dominance based rough sets in the search for genes under balancing selection. Transactions on Rough Sets XI, Lecture Notes in Computer Science 5946, p. 53÷65.
73. Cyran K. A., Kimmel M. (2010): Alternatives to the Wright-Fisher model: The robustness of the mitochondrial Eve dating. Theor. Pop. Biol. 78(3), p. 165÷172.
74. Czech Z. J. (2010): Wprowadzenie do obliczeń równoległych. PWN, Warszawa.
75. Danoeux T. (1997): Neural Network Applications – Pattern classification. In In Fiesler E., Beale R. (eds.) Handbook of Neural Computation, IOP Publishing Ltd and Oxford University Press, p. F1.2:1÷F1.2:8.
76. Darwin Ch. (1859): On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.
77. De Jong K., Fogel D. B., Schwefel H. P. (1997): A history of evolutionary computation. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. A2.3:1÷A2.3:12.
78. Demetrius L., Schuster P., Sigmund K. (1985): Polynucleotide Evolution and Branching Processes. Bull. Math. Biol. 47, p. 239÷262.

79. Dempster A. P., Laird N. M., Rubin D. B. (1977): Maximum likelihood from incomplete data via the EM algorithm. With discussion. J. Roy. Stat. Soc. Ser. B 39, p. 1÷38.
80. Doherty P, Szałas A. (2004): On the correspondence between approximations and similarity. Lecture Notes in Artificial Intelligence 3066, p. 143÷152.
81. Donnelly M. J., Licht M. C., Lehmann T. (2001): Evidence for recent population expansion in the evolutionary history of the malaria vectors Anopheles arabiensis and Anopheles gambiae. Mol. Biol. Evol. 18(7), p. 1353÷1364.
82. Draper W. E., Hayden E. J., Lehman N. (2008): Mechanisms of covalent self-assembly of the Azoarcus ribozyme from four fragment oligonucleotides. Nucleic Acids Res. 36, p. 520÷31.
83. Dreyfus H. (1972): What Computers Can't Do. MIT Press, New York.
84. Dyson F. (1999): The origins of Life. Cambridge University Press.
85. Edwards M. R. (1998): From a soup or a seed? Pyritic metabolic complexes in the origin of life. Trends Ecol. Evol. 13, p. 179÷181.
86. Eigen M., Gardiner W., Schuster P., Winckler-Oswatitch R. (1981): The Origin of Genetic Information. Sci. Am. 244(4), p. 88÷118.
87. Eigen M., Schuster P. (1977): The Hypercycle – A Principle of Natural Self-Organization. Naturwissenschaften 64(11), p. 541÷565.
88. Ellis N. A., Roe A. M., Kozloski J., Proytcheva M., Falk C., German J. (1994): Linkage disequilibrium between the FES, D15S127, and BLM loci in Ashkenazi Jews with Bloom syndrome. Am. J. Hum. Genet. 55, p. 453÷460,
89. Evans P.D., Anderson J. R., Vallender E. J., Gilbert S. L., Malcom Ch. M., et al. (2004): Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. Human Molecular Genetics 13, p. 489÷494.
90. Ewens W. J. (2003): Mathematical population genetics. Second edition. Springer-Verlag, New York.
91. Excoffier L., Slatkin M. (1995): Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12, p. 921÷927.
92. Eyre-Walker A., Awadalla Ph. (2001): Does Human mtDNA Recombine? J. Mol. Evol. 53, p. 430÷435.
93. Fares A., Bouzid A., Hamdi M. (2000): Rotation invariance using diffraction pattern sampling in optical pattern recognition. J. of Microwaves and Optoelect. 2(2), p. 33÷39.
94. Feller W. (1968): An Introduction to Probability and Its Applications. Vol. 1, 3rd ed., Wiley, New York.
95. Ferris J. P. (2002): Montmorillonite catalysis of 30–50 mer oligonucleotides: laboratory demonstration of potential steps in the origin of the RNA world. Orig. Life Evol. Biosph. 32, p. 311÷332.

96. Ferris J. P. (2006): Montmorillonite-catalysed formation of RNA oligomers: The possible role of catalysis in the origins of life. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361, p. 1777÷1786.
97. Ferris J. P., Ertem G. (1993): Montmorillonite catalysis of RNA oligomer formation in aqueous solution. A model for the prebiotic formation of RNA. J. Am. Chem. Soc. 115, p. 1227÷12275.
98. Ferris J. P., Hill A. R., Liu R., Orgel L. E. (1996) Synthesis of long prebiotic oligomers on mineral surfaces. Nature 381, p. 59÷61.
99. Fisher M. C., Koenig G. L., White Th. J., San-Blas G., Negroni R., Gutiérrez Alvarez I., Wanke B., Taylor J. W. (2001): Biogeographic range expansion into South America by Coccidioides immitis mirrors New World patterns of human migration. Proc. Natl. Acad. Sci. USA 98(8), p. 4558÷4562.
100. Fogel D. B. (1997a): Why evolutionary computation – Introduction. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. A1.1:1÷A1.1:2.
101. Fogel D. B. (1997b): Evolutionary Computation Models – Search Operators – Mutation – Real-valued vectors. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C3.2:2÷C3.2:5.
102. Fogel D. B. (1997c): Evolutionary Computation Models – Search Operators – Recombination – Real-valued vectors. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C3.3:11÷C3.3:13.
103. Fogel D. B. (1997d): Evolutionary Computation Models – Representations – Finite-state representations. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C1.5:1÷C1.5:3.
104. Fogel L. J., Owens A. J., Walsh M. J. (1966): Artificial intelligence through simulated evolution. Wiley, New York.
105. Fonseca C. M., Fleming P. J. (1997): Evolutionary Computation Models – Fitness Evaluation – Multiobjective optimization. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C4.5:1÷C4.5:9.
106. Forster P. (2004): Ice Ages and miochondrial DNA chronology of human dispersals: a review. Phil. Trans. R. Soc. Lond. B. 359, p. 255÷264.
107. Franklin S. (1997): Artificial Minds, MIT Press, Cambridge Massachusetts.

108. Frayer D.W. (1986): Cranial variation at Mladeč and relationship between Mousterian and Upper Paleolithic hominidy. Anthropos 23, p. 243÷256.
109. Frayer D. W. (1992): Evolution at the European edge: Neanderthal and Upper Paleolithic relationships. Prehistoric Europeenne 2, p. 9÷69.
110. Fu Y. X., Li W. H. (1993): Statistical Tests of Neutrality of Mutations. Genetics 133, p. 693÷709.
111. Fu Y. X. (1996): New Statistical Tests of Neutrality for DNA Samples From a Population. Genetics 143, p. 557÷570.
112. Fu Y. X. (1997): Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics 147, p. 915÷925.
113. Fu Y. X. (2003): Population genetics – Course outline. Materials for the PhD students course in MD Anderson Cancer Center, Houston.
114. Fullerton S. M., Harding R., Boyce A., Clegg J. (1994): Molecular and population genetic analysis of allelic sequence diversity at the human beta-globin locus. Proc. Nat. Acad. Sci. USA 91, p. 1805÷1809.
115. Ganotra D., Joseph J., Singh K. (2002): Neural network based face recognition by using diffraction pattern sampling with a digital ring-wedge detector. Opt Comm. 202, p. 61÷68.
116. Ganotra D., Joseph J., Singh K. (2003): Modified geometry of ring-wedge detector for sampling Fourier transform of fingerprints for classification using neural networks. Proc SPIE 4829, p. 407÷408.
117. George N., Wang S. (1994): Neural networks applied to diffraction-pattern sampling. Appl. Opt. 33, p. 3127÷3134.
118. Gilad Y., Rosenberg S., Przeworski M., Lancet D., Skorecki K. (2002): Evidence for positive selection and population structure at the human MAO-A gene. Proc. Natl. Acad. Sci. 99, p. 862÷867.
119. Gilbert W. (1986): The RNA World. Nature 319, p. 618÷618.
120. Gillespie J. H. (1998): Population Genetics – A Concise Guide. The John Hopkins University Press, Baltimore and London.
121. Gödel K. (1931): Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme I. Monatshefte für Mathematik und Physik 37, p. 173÷198.
122. Goldberg D. E., Deb K. (1991): A comparative analysis of selection schemes used in genetic algorithms. In Rawlins G., San Mateo C. A. (eds.) Foundations of Genetic Algorithms, p. 69÷93.
123. Goldberg D. E. (1989): Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company, Inc. Massachusetts.
124. Goldstein D. B., Pollock D.D. (1997): Launching Microsatellites: A Review of Mutation Processes and Methods of Phylogenetic Inference. J. Hered. 88(5), p. 335÷42.

125. Gomolińska A. (2002): A comparative study of some generalized rough approximations. Fundamenta Informaticae 51(1), p. 103÷119.
126. Greco S., Matarazzo B., Slowinski R. (1998): A new rough set approach to evaluation of bankruptcy risk. In Zopounidis C. (ed.) Operational Tools in the Management of Financial Risk, Dordrecht, Boston: Kluwer Academic Publishers, p. 121÷136.
127. Greco S., Matarazzo B., Słowinski R. (1999a): Rough Approximation of Preference Relation by Dominance Relations. European Journal of Operational Research 117, p. 63÷83.
128. Greco S., Matarazzo B., Slowinski R. (1999b): The use of rough sets and fuzzy sets in MCDM. In Gal T., Hanne T., Stewart T. (Eds.) Advances in Multiple Criteria Decision Making, Dordrecht, Boston: Kluwer Academic Publishers, p. 14.1÷14.59.
129. Greco S., Matarazzo B., Słowinski R., Stefanowski J. (2001): Variable Consistency Model of Dominance-based Rough Sets Approach, Lecture Notes in Computer Science. 2005, p. 170÷181.
130. Green R. E., Krause J., Ptak S. E., Briggs A. W., Ronan M. T., Simons J. F., Du L., Egholm M., Rothberg J. M., Paunovic M., Pääbo S. (2006): Analysis of one million base pairs of Neanderthal DNA. Nature 444, p. 330÷336.
131. Green R. E., Malaspinas A.-S., Krause J., Briggs A. W., Johnson Ph. L. F., Uhler C., Meyer M., Good J. M., Maricic T., Stenzel U., Pruefer K., Siebauer M., Burbano H. A., Ronan M., Rothberg J. M., Egholm M., Rudan P., Brajkovic D., Kucan Z., Gusic I., Wikstrom M., Laakkonen L., Kelso J., Slatkin M., Pääbo S. (2008): A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. Cell 134, p. 416÷426.
132. Green R. E., Krause J., Briggs A. W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M. H-Y, Hansen N. F., Durand E. Y., Malaspinas A.-S., Jensen J. D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H. A., Good J. M., Schultz R., Aximu-Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum Ch., Lander E. S., Russ C., Novod N., Affourtit J., Egholm M., Verna Ch., Rudan P., Brajkovic D., Kucan Ž., Gušic I., Doronichev V. B., Golovanova L. V., Lalueza-Fox C., de la Rasilla M., Fortea J., Rosas A., Schmitz R. W., Johnson Ph. L. F., Eichler E. E., Falush D., Birney E., Mullikin J. C, Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S. (2010): A draft sequence of the Neandertal genome. Science 328, p. 710÷721.
133. Grefenstette J. (1997a): Evolutionary Computation Models – Selection – Proportional selection and sampling algorithms. In Bäck T., Fogel D. B., Michalewicz Z. (eds.): Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C2.2:1÷C2.2:7.

134. Grefenstette J. (1997b): Evolutionary Computation Models – Selection – Rank-based selection. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C2.4:1÷C2.4:6.
135. Griffiths R. C., Tavare S. (1995): Unrooted genealogical tree probabilities in the infinitely-many-sites model. Math. Biosci. 127, p. 77÷98.
136. Grzymała-Busse J. W. (2003): Rough set strategies to data with missing attribute values. Proceedings of the Workshop on Foundations and New Directions in Data Mining, associated with the third IEEE International Conference on Data Mining, Melbourne, FL, USA, p. 56÷63.
137. Grzymała-Busse J. W. (2004): Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Lecture Notes in Computer Science 3100, p. 78÷95.
138. Guyon I., Gunn S., Nikravesh M., Zadeh L. A. (2006): Feature extraction. Foundations and applications. Berlin, Springer.
139. Hameroff S. R., Penrose R. (1996) Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. In Hameroff S. R., Kaszniak A., Scott A. C. (eds.) Toward a Science of Consciousness - The First Tucson Discussions and Debates, MIT Press, Cambridge, MA.
140. Harding R. M., Fullerton S. M., Griffiths R. C., Bond J., Cox M. J., Schneider J. A., Moulin D., Clegg J. B. (1997): Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. 60, p. 772÷798.
141. Hartl D. L., Clark A. G. (1997): Principles of Population Genetics. Sinauer Assoc., Sunderland, MA.
142. Hasegawa M., Horai S. (1991): Time of the deepest root for polymorphism in human mitochondrial DNA. J. Mol. Evol. 32(1), p. 37÷42.
143. Hebb D. O. (1949): The organization of behavior. Wiley, New York.
144. Hein J., Schierup M. H., Wiuf C. (2005): Gene genealogies, variation and evolution: a primer in coalescent theory. Oxford, New York, Oxford University Press.
145. Hertz J., Krogh A., Palmer R. G. (1991): Introduction to the theory of neural computation. Addison-Wesley Publishing Company, Redwood City.
146. Hey J. (1997): Mitochondrial and nuclear gene trees present conflicting portraits of human origins. Mol. Biol. Evol. 14, p. 166÷172.
147. Hoogs A., Collins R., Kaucic R., Mundy J. (2003): A common set of perceptual observables for grouping, figure-ground discrimination, and texture classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(4), p. 458÷474.
148. Holland J. H. (1967): Nonlinear environments permitting efficient adaptation. Computer and Information Sciences II, Academic, New York.

149. Hopfield J.J. (1982): Neural networks and physical systems with emergent collective computational abilities. Proc. Nat. Acad. Sci. USA 79, p. 2554÷2558.
150. Hoyle F., Wickramasinghe N.C. (1999): Astronomical Origins of Life – Steps Towards Panspermia. Kluwer Academic Publishers.
151. Huang S., Li B. B., Gray M. D., Oshima J., Mian I. S., Campisi J. (1998): The premature ageing syndrome protein, WRN, is a 3-prime-5-prime exonuclease. Nature Genet. 20, p. 114÷115.
152. Huang W., Ferris J. P. (2006): One-step, regioselective synthesis of up to 50-mers of RNA oligomers by montmorillonite catalysis. J. Am. Chem. Soc. 2006, p. 8914÷8919.
153. Hudson R. R. (1987): Estimating the recombination parameter of a finite population model without selection. Genet. Res. 50, p. 245÷250.
154. Hudson R. R, Kreitman M., Aguade M. (1987): A test of neutral molecular evolution based on nucleotide data. Genetics 116, p. 153÷159.
155. Izhikevich E. M., Edelman G. M. (2008): Large-scale model of mammalian thalamocortical systems. Proc. Natl. Acad. Sci. USA 105(9), p. 3593÷3598.
156. Jaroszewicz L. R, Cyran K. A., Podeszwa T. (2000): Optimized CGH-based pattern recognizer. Opt Appl. 30, p. 317÷333.
157. Jaroszewicz L. R, Merta I., Podeszwa T., Cyran K. A. (2002): Airplane engine condition monitoring system based on artificial neural network. In McNulty G. J. (ed.) Quality, Reliability and Maintenance, Bury St. Edmunts London, Professional Engineering Publishing, p. 179÷182.
158. Järvinen J. (2001): Approximations and roughs sets based on tolerances. Lecture Notes in Artificial Intelligence 2005, p. 182÷189.
159. Jobling M. (2001): In the name of the father: surnames and genetics. Trends in Genetics 17, p. 353÷357.
160. Jobling M. A., Hurles M. E., Tyler-Smith C. (2004): Human Evolutionary Genetics: origins, peoples & disease. Garland Science, New Delhi, India.
161. Johnston W. K., Unrau P. J., Lawrence M. S., Glasner M. E., Bartel D. P. (2001): RNA-catalyzed RNA polymerization: Accurate and general RNA-template primer extension. Science 292, p. 1319÷1325.
162. Joyce G. F. (1989): RNA evolution and the origins of life. Nature 338, p. 217÷224.
163. Joyce G. F. (2005): Evolution in an RNA world. Orig. Life Evol. B 36, p. 202÷204.
164. Joyce G. F., Orgel L. E. (2006): Progress toward Understanding the Origin of the RNA World. In Gesteland R. F. Cech Th. R., Atkins J. F. (eds.) The RNA World – Third Edition, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

165. Jutten C. (1997): Supervised composite networks. In Fiesler E., Beale R. (eds.) Handbook of Neural Computation, IOP Publishing and Oxford University Press, Philadelphia, New York, Oxford, C1.6.1÷C.1.6.13.
166. Karow J. K., Constantinou A., Li J.-L., West S. C., Hickson I. D. (2000): The Bloom's syndrome gene product promotes branch migration of Holliday junctions. Proc. Nat. Acad. Sci. 97, p. 6504÷6508.
167. Kaye P. H., Barton J. E., Hirst E., Clark J. M. (2000): Simultaneous light scattering and intrinsic fluorescence measurement for the classification of airbone particles. App. Opt. 39(21), p. 3738÷3745.
168. Kelly J. K. (1997): A test of Neutrality Based on Interlocus Associations. Genetics 146, p. 1197÷1206.
169. Khanna K. K., Keating K. E., Kozlov S., Scott S., Gatei M., et al. (1998): ATM associates with and phosphorylates p53: mapping the region of interaction. Natur. Genet. 20, p. 398÷400.
170. Kimmel M., Chakraborty R., King J., Bamshad M., Watkins W., Jorde L. (1998): Signatures of population expansion in microsatellite repeat data. Genetics 148, p. 1921÷1930.
171. Kimmel M., Axelrod D. E. (2002): Branching Processes in Biology. New-York: Springer-Verlag.
172. Kimura M., Ohta T. (1971): Protein polymorphism as a phase of molecular evolution. Nature 229, p. 467÷469.
173. Kimura M., Ohta T. (1978): Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. USA 75(6), p. 2868÷2872.
174. Kimura M. (1983): The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.
175. King J. P., Kimmel M., Chakraborty R. (2000): A power analysis microstallite-based statistics for inferring past population growth. Mol. Biol. Evol. 17(12), p. 1859÷1868.
176. Klebaner F. C., Sagitov S. (2002): The age of a Galton-Watson population with a geometric offspring distribution. J. Appl. Prob. 39, p. 816÷828.
177. Kohonen T. (1984): Self-organization and associative memory. Springer Verlag, Heidelberg 1984.
178. Kohonen T. (1990): The self-organizing map. Proc. IEEE, Special Issue on Neural Networks 78(9), p. 1464÷1480.
179. Korbicz J., Obuchowicz A., Uciński D. (1994): Sztuczne sieci neuronowe – podstawy i zastosowania. Akademicka Oficyna Wydawnicza PLJ, Warszawa, Poland.
180. Kreis T. (1996): Holographic interferometry: Principles and methods. Akademie Verlag Series in Optical Metrology. Vol 1. Akademie-Verlag, Berlin.

181. Krings M., Stone A., Schmitz R., Krainitzki H., Stoneking M., Pääbo S. (1997): Neandertal DNA sequences and the origin of modern humans. Cell 90, p. 19÷30.
182. Krings M., Geisert H., Schmitz R., Krainitzki H., Pääbo S. (1999): DNA sequence of the mitochondrial hypervariable region II from the Neandertal type specimen. Proc. Natl. Acad. Sci. USA 96, p. 5581÷5585.
183. Krings M., Capelli C., Tschentscher F., Geisert H., Meyer S., von Haeseler A., Grossschmidt K., Possnert G., Paunovic M., Pääbo S. (2000) A view of Neandertal genetic diversity. Nature Genetics 26, p. 144÷146.
184. Kurzweil R. (2005): The Singularity is Near. Viking Press, New York.
185. Laan M., Wiebe V., Khusnutdinova E., Remm M., Pääbo S. (2005): X-chromosome as a marker for population history: linkage disequilibrium and haplotype study in Euroasians populations. Eur. J. Hum. Genet. 13(4), p. 452÷462.
186. Lawrence J. (1994): Introduction to Neural Networks, California Scientific Software Press, Nevada City.
187. Lambert A. (2003): Coalescence times for the branching process. Adv. Appl. Prob. 35, p.1071÷1098.
188. Leakey M., Walker A. (2003): Early hominid fossils from Africa. Scientific American, Special edition: New look at human evolution, 14÷19.
189. Lerner V. (2003): Variation Principle in Informational Macrodynamics. Kluwer Academic Publishers, Boston, Dordrecht, London.
190. Li A., Swift M. (2000): Mutations at the ataxia-telangiectasia locus and clinical phenotypes of A-T patients. Am. J. Med. Genet. 92, p. 170÷177.
191. Lim D.-S., Kirsch D. G., Canman C. E., Ahn J.-H., Ziv Y., et al. (1998): ATM binds to beta-adaptin in cytoplasmic vesicles. Proc. Natl. Acad. Sci. USA 95, p. 10146÷10151.
192. Lutay A. V., Zenkova M. A., Vlassov V. V. (2007): Nonenzymatic Recombination of RNA: Possible Mechanism for the Formation of Novel Sequences. Chem. & Biod. 4, p. 762 ÷ 767.
193. Łęski J. (2008): Systemy neuronowo-rozmyte. WNT, Warszawa.
194. Ma W. T., Yu C.W. (2006): Intramolecular RNA replicase: Possibly the first self-replicating molecule in the RNA world. Orig. Life Evol. Biosph. 36, p. 413÷420.
195. Ma W., Yu Ch., Zhang W. (2007a): Monte Carlo simulations of early molecular evolution in the RNA World. Biosystems 90, p. 28÷39.
196. Ma W., Yu C., Zhang W., Hu J. (2007b): Nucleotide synthetase ribozymes may have emerged first in the RNA world. RNA 13, p. 2012÷2019.
197. Mahfoud S. W. (1997): Evolutionary Computation Models – Selection – Boltzmann selection. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C2.5:1÷C2.5:4.

198. Mait J. N, Athale R., van der Gracht J. (2003): Evolutionary paths in imaging and recent trends. Optics Express 11(18), p. 2093÷2101.
199. Manapat M., Ohtsuki H., Bürger R., Nowak M. A. (2009): Originator dynamics. J. Theor. Biol. 256, p. 586÷595.
200. Marjoram P., Wall J. D. (2006): Fast "coalescent" simulation. BMC Genet. 7(16), p. doi:10.1186/1471-2156-7-16.
201. Marsaglia G., Zaman A., Tsang W. W. (1990): Toward a universal random number generator. Statist. Prob. Lett 8, p. 35÷39.
202. Marsaglia G. (1993): Monkey tests for random number generators. Comput. Math. Appl. 9, p. 1÷10.
203. Matsumoto M., Nishimuram T. (1998): Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM TOMACS 8, p. 3÷30.
204. McCulloch W. S., Pitts W. (1943): A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, p. 115÷133.
205. McDonald J. H., Kreitman M. (1991): Adaptive protein evolution at the Adh locus in Drosophila. Nature 351, p. 652÷654.
206. McGinness K., Joyce G. F. (2003): In search of an RNA Replicase Ribozyme. Chemistry & Biology 10, p. 5÷14.
207. McVean G. (2002): Natural Selection. Printed Materials of Univ. Oxford, Dept. Stat., p. 1÷25.
208. Mellars P. (2004): Neanderthals and the modern human colonization of Europe. Nature 432, p. 461÷465.
209. Merleau-Ponty M. (1945): Phenomenology of perception. Paris and New York: Smith, Gallimard, Paris and Routledge & Kegan Paul. trans. by Colin Smith.
210. Michalewicz Z. (1992): Genetic Algorithms + Data Structures = Evolution Programs. Springer – Verlag, Berlin – Heidelberg.
211. Minsky M., Papert S. (1969): Perceptrons. MIT Press, Cambridge.
212. Mirazon Lahr M., Foley R. (2004): Human evolution writ small. Nature 431, p. 1043÷1044.
213. Monnard P. A., Szostak J. W. (2008): Metal-ion catalyzed polymerization in the eutectic phase in water-ice: A possible approach to template-directed RNA polymerization. J. Inorg. Biochem. 102, p. 1104÷1111.
214. Moravec H. (1988): Mind Children: The Future of Robot and Human Intelligence. Harvard University Press, Cambridge, Massachusetts, London.

215. Morwood M. J., Soejono R. P., Roberts R. G., Sutikna T., Turney C. S. M., Wesaway K. E., Rink W. J., Zhao J.-X., Van den Bergh G. D., Rokus Awe Due, Hobbis D. R., Moore M. W., Bird M. I., Fifield L. K. (2004): Archeology and age of a new hominin from Flores in eastern Indonesia. Nature 431, p. 1087÷1091.
216. Moya-Sola S., Köhler M., Alba D. M., Casanovas-Vilar I., Galindo J. (2004): Pierolapithecus catalaunicus a new middle Miocene Great Ape from Spain. Science 306, p. 1339÷1344.
217. Mrózek A. (1992a): Rough sets in computer implementation of rule-based control of industrial processes. In Słowiński R. (ed.) Intelligent decision support. handbook of applications and advances of the rough sets, Kluwer Academic Publishers, Boston, London Dordrecht, p. 19÷31.
218. Mrózek A. (1992b): A new method for discovering rules from examples in expert systems. Man-Machine Studies 36, p. 127÷143.
219. Mrózek A., Plonka L. (1993): Rough sets in image analysis. Foundations of Computing and Decision Sciences 18(3-4), p. 268÷273.
220. Mrózek A. (1998): Rough sets – personal communication.
221. Myszor D., Cyran K. A. (2009): Estimation of the number of primordial genes in compartment model of RNA World. In Cyran K.A. et al. (eds.) Advances in Intelligent and Soft Computing, Springer, 59, p. 151÷161.
222. Myszor D., K. A. Cyran (2010): Influence of non-enzymatic template-directed RNA recombination processes on polynucleotides lengths in Monte Carlo simulation model of the RNA World. Int. J. Appl. Math. & Informatics 1(4), p. 1÷8.
223. Nagylaki T. (1990): Models and approximations for random genetic drift. Theor. Popul. Biol. 37, p. 192÷212.
224. Nebeker B. M., Hirleman E. D. (2000): Light scattering by particles and defects on surfaces: semiconductor wafer inspector. Lecture Notes in Physics 534, p. 237÷257.
225. Nechaev S. Y., Lutay A. V., Vlassov V. V., Zenkova M. A. (2009): Non-Enzymatic Template-Directed Recombination of RNAs. Int. J. Mol. Sci. 10, p. 1788÷1807.
226. Nielsen R., Weinreich D. M. (1999): The Age of Nonsynonymous and Synonymous Mutations and Implications for the Slightly Deleterious Theory. Genetics 153, p. 497÷506.
227. Nielsen R. (2001): Statistical tests of selective neutrality in the age of genomics. Heredity 86, p. 641÷647.
228. Niesert U., Harnasch D., Bresch C. (1981): Origin of life - between Scylla and Charybdis. J. Mol. Evol. 17(6), p. 348÷53.
229. Niesert U. (1987): How many genes to start with? A computer simulation about the origin of life. Orig. Life Evol. Biosph. 17(2), p. 155÷69.

230. Noonan J. P., Coop G., Kudaravalli S., Smith D., Krause J., Alessi J., Chen F., Platt D., Pääbo S., Pritchard J. K, Rubin E. M. (2006): Sequencing and analysis of Neanderthal genomic DNA. Science 314, p. 1113÷1118.
231. Nowak M. A., Ohtsuki H. (2008): Prevolutionary dynamics and the origin of evolution. Proc. Natl. Acad. Sci. USA. 105, p. 14924÷14927.
232. O'Connell N. (1995): The genealogy of branching processes and the age of our most recent common ancestor. Adv. Appl. Prob. 27, p. 418÷442.
233. Ohtsuki H., Nowak M. A. (2009) Prelife catalysts and replicators. Proc. R. Soc. B 276, p. 3783÷3790.
234. Orgel L. E. (1998): The origin of life – a review of facts and speculations. Trends Biochem. Sci. 23(12), p. 491÷495.
235. Orgel L. E. (2004): Prebiotic chemistry and the origin of the RNA world. Crit. Rev. Biochem. Mol. Biol. 39, p. 99÷123.
236. Osowski S. (1996): Sieci neuronowe w ujęciu algorytmicznym. WNT, Warszawa.
237. Ovchinnikov I., Götherström A., Romanova G., Kharitonov V., Lidén K., Goodwin W. (2000): Molecular analysis of Neanderthal DNA from the northern Caucasus. Nature 404, p. 490÷493.
238. Pal S. K., Peters J. F. (2010): Rough Fuzzy Image Analysis. Chapman & Hall/CRC Press, Mathematical and Computational Imaging Science Series.
239. Paun G., Rozenberg G., Salomaa A. (1998): DNA Computing – New Computing Paradigm. Springer, Berlin, Heidelberg.
240. Pavel M. (1993): Fundamentals of pattern recognition. 2nd ed. Marcel Dekker, Inc., N.Y., U.S.A.
241. Pawlak Z. (1982): Rough sets. International Journal of Information and Computer Sciences 11, p. 341÷356.
242. Pawlak Z. (1991): Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston, London, Dordrecht.
243. Pawlak Z. (1995a): Rough Sets Rudiments. Report supported by State Committee for Scientific Research in grant No. 8S503 021 06, Warszawa.
244. Pawlak Z. (1995b): Wiedza a zbiory przybliżone. Podstawowe problemy współczesnej techniki, Problemy sztucznej Inteligencji, Tom XXVIII, Wiedza i Życie, Warszawa, p. 9÷21.
245. Pawlak Z., Skowron A. (2007a): Rough sets and Boolean reasoning. Information Sciences 177, p. 41÷73.
246. Pawlak Z., Skowron A. (2007b): Rough sets: Some extensions. Information Sciences 177, p. 28÷40.

247. Pawlak Z., Skowron A. (2007c): Rudiments of rough sets. Information Sciences 177, p. 3÷27.
248. Pennisi E. (2006): The dawn of the stone age genomics. Science 314, p. 1068÷1071.
249. Pennisi E. (2007): No sex please, We're Neandertals. Science 318, p. 967-967.
250. Penrose R. (1989): The Emperor's New Mind. Oxford University Press, Oxford.
251. Peters J. F. (2007): Near sets. General theory about nearness of objects. Applied Mathematical Sciences 1(53), p. 2609÷2029.
252. Peters J. F. (2009): Discovering affinities between perceptual granules: L2 norm-based tolerance near preclass approach. In: Cyran K.A. et al. (eds.) Advances in Intelligent and Soft Computing, Springer, 59, p. 43÷55.
253. Peters J. F., Skowron A., Stepaniuk J. (2007): Nearness of objects: Extension of approximation space model. Fundamenta Informaticae 79(3-4), p. 497÷512.
254. Peters J. F., Ramanna S. (2009): Affinities between perceptual granules: Foundations and perspectives. In Bargiela A., Pedrycz W. (eds.) Human-centric information processing through granular modelling, Berlin: Springer-Verlag 182, p. 49÷66.
255. Peters, J. F., Wasilewski P. (2009): Foundations of near sets. Information Sciences 179, p. 3091÷3109.
256. Piekara A. H. (1976): Nowe aspekty optyki – wstęp do elektroniki kwantowej i w szczególności do optyki nieliniowej i optyki światła spójnego. PWN, Warsaw.
257. Plagnol V., Wall J. D. (2006): Possible ancestral structure in human populations. PloS Genetics 2(7), p. 0972÷0979.
258. Podeszwa T., Jaroszewicz L. R, Cyran K. A. (2003): Fiberscope based engine condition monitoring system. Proc SPIE 5124, p. 299÷303.
259. Polańska J. (2003): The EM algorithm and its implementation for the estimation of the frequencies of SNP-haplotypes. Int. J. Appl. Math. Comput. Sci. 13, p. 419÷429.
260. Polański A., Kimmel M., Chakraborty R. (1998): Application of time-dependent coalescence process for inferring the history of population size changes from DNA sequence data. Proc. Natl. Acad. Sci. USA. 95, p. 5456÷5461.
261. Polański A., Kimmel M. (2003): Population genetics models for the statistics of DNA samples under different demographic scenarios – maximum likelihood versus approximate methods. Int. J. Appl. Math. Comput. Sci. 13, p. 347÷355.
262. Pribram K. H. (1991): Brain and Perception: Holonomy and Structure in Figural Processing. Lawrence Erlbaum Associates, Inc. Publishers, New Jersey.
263. Pross A. (2004): Causation and the Origin of Life. Metabolism or Replication First? Origins of Life and Evolution of the Biosphere 34(3), p. 307÷321.

264. Radcliffe N. J. (1997): Schema processing. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. B2.5:1÷B2.5:10.
265. Raghu P. P., Yegnanrayana B. (1998): Supervised texture classification using a probabilistic neural network and constraint satisfaction model. IEEE Trans Neural Networks 9, p. 516÷522.
266. Reich D. E., Feldman M. W., Goldstein D. B. (1999): Statistical Properties of Two Tests that Use Multilocus Data Sets to Detect Population Expansions. Mol. Biol. Evol. 16, p. 453÷466.
267. Reich D. E., Goldstein D. B. (1998): Genetic evidence for a Paleolithic human population expansion in Africa. Proc. Natl. Acad. Sci. USA 95, p. 8119÷8123.
268. Relethford J. H. (2001): Absence of regional affinities of Neandertal DNA with living humans does not reject multiregional evolution. Am. J. Phys. Anthropology 115, p. 95÷98.
269. Renwick A., Davison L., Spratt H., King J. P., Kimmel M. (2001): DNA Dinucleotide Evolution in Humans: Fitting Theory to Facts. Genetics 159(2), p. 737÷747.
270. Rode B. M., Fitz D., Jakschitz T. (2007): The first steps of chemical evolution towards origin of life. Chemistry & Biodiversity 4, p. 2674÷2702.
271. Rogers A. (1995): Genetic evidence for Pleistocene population explosion. Evolution 49, p. 608÷615.
272. Rosenblatt F. (1958): The perceptron: a probabilistic model for information storage and organization in the brain. Psych. Rev. 65, p. 386÷408.
273. Rudolf G. (1997): Stochastic processes. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. B2.2:1÷B2.2:8
274. Rumelhart D. E., Hinton G. E., Williams R. J. (1986a): Learning representations by back-propagating errors. Nature 323, p. 533÷536.
275. Rumelhart D. E., Hinton G. E., Williams R. J. (1986b): Learning internal representations by error propagation. In Rumelhart D. E., McClelland J. L. (eds.) Parallel Distributed Processing, p. 318÷362.
276. Rumelhart D. E., Durbin R., Golden R., Chauvin Y. (1992): Backpropagation: Theoretical foundations. In Chauvin Y., Rumelhart D. E. (eds.) Backpropagation and Connectionist Theory, Lawrence Erlbaum.
277. Russell S. J., Norvig P. (2003): Artificial Intelligence: A Modern Approach (2nd ed.). Upper Saddle River, New Jersey: Prentice Hall.
278. Rutkowska D., Piliński M., Rutkowski L. (1999): Sieci neuronowe, algorytmy genetyczne i systemy rozmyte. Wydawnictwo Naukowe PWN, Łódź, Poland.

279. Sagre D., Lancet D. (1999): A Statistical Chemistry Approach to the Origin of Life. Chemtracts–Biochem. Mol. Biol. 12(6), p. 382÷397.
280. Sarma J., De Jong K. (1997): Evolutionary Computation Models – Selection – Generation gap methods. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C2.2:1÷C2.2:7.
281. Schmitz R., Bonani G., Smith F. H. (2002): New research at the Neandertal type site in the Neander Valley of Germany. Journal of Human Evolution 42, p. A32÷A32.
282. Schwefel H. P. (1965): Kybernetische Evolution als Strategie der Experimentallen Forschung in der Strömungstechnik, Diploma Thesis, Technical University of Berlin.
283. Searle J. (1980)Ł Minds, Brains and Programs. Behavioral and Brain Sciences 3(3), p. 417÷457.
284. Searle J. (1999): Mind, language and society. NY: Basic Books, New York.
285. Serre D., Langaney A., Chech M., Teschler-Nicola M., Paunovic M., Mennecier P., Hofreiter M., Possnert G., Pääbo S. (2004): No evidence of Neanderthal mtDNA contribution to early modern humans. PLOS Biology (2), p. 313÷317.
286. Sia E. A., Butler Ch. A., Dominska M., Greenwell P., Fox Th. D., Petes Th. D. (2000): Analysis of microsatellite mutations in the mitochondrial DNA of Saccharomyces cerevisiae. Proc. Natl. Acad. Sci. USA 97(1), p. 250÷255.
287. Siitonen H. A., Kopra O., Haravuori H., Winter R. M., Saamanen A. M., et al. (2003): Molecular defect of RAPADILINO syndrom expands the phenotype spectrum of RECQL diseases. Hum. Mol. Genet. 12(21), p. 2837÷2844.
288. Sinclair D. A., Mills K., Guarente L. (1997): Accelerated aging and nucleolar fragmentation in yeast sgs1 mutants. Science 277, p. 1313÷1316.
289. Skowron A., Rauszer C. (1992): The discernibility matrices and functions in information systems. In Słowiński R. (ed) Intelligent Decision Support. Handbook of Applications and Advances of Rough Set Theory, Dordrecht, Kluwer Academic Publishers, p. 311÷362.
290. Skowron A., Grzymała-Busse J. W. (1994): From rough set theory to evidence theory. In Yager R. R et al. (eds) Advances in Dempster Shafer Theory of Evidence, New York, Wiley & Sons, p. 193÷236.
291. Skowron A, Stepaniuk J. (1996): Tolerance approximation spaces. Fundamenta Informaticae 27, p. 245÷253.
292. Słowiński R, Vanderpooten D. (1997): Similarity relation as a basis for rough approximations. In: Wang P. P. (ed) Advances in Machine Intelligence and Soft Computing, Bookwrights, Raleigh, p. 17÷33.
293. Słowiński R, Vanderpooten D. (2000): A generalized definition of rough approximations based on similarity. IEEE Transaction on Data and Knowledge Engineering 12(2), p. 331÷336.

294. Slatkin M., Rannala B. (2000): Estimating allele age. Annual Review of Genomics and Human Genetics 1, p. 225÷249.
295. Smith F. H. (1984): Fossils hominids from the Upper Pleistocene of Central Europe and the origin of modern Europeans. In Spencer F. (ed.) The origins of modern humans: A world survey of the fossil evidence, New York, p. 137÷210.
296. Smith J. M., Szathmary E. (1999): The Origins of Life. From the Birth of Life to the Origin of Language. Oxford University Press, 1999.
297. Stanczyk U., Cyran K. A. (2007a): Machine learning approach to authorship attribution of literary texts. International Journal of Applied Mathematics and Informatics 1(4), p. 151÷158.
298. Stanczyk U., Cyran K. A. (2007b): On employing elements of Rough Set Theory to stylometric analysis of literary texts. International Journal of Applied Mathematics and Informatics 1(4), p. 159÷166.
299. Stanczyk U., Cyran K. A., Pochopien B. (2007): Theory of logic circuits: vol.2, Circuit design and analysis. Publishers of the Silesian University of Technology, Gliwice.
300. Steitz T. A., Moore P. B. (2003): RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem. Sci. 28, p. 411÷418.
301. Sun R. (2002): Duality of the Mind. Lawrence Erlbaum Associates, Mahwah, NY.
302. Szathmary E., Demeter L. (1987): Group Selection of Early Replicators and the Origin of Life. J. Theor. Biol. 128, p. 463÷486.
303. Szostak J. W. (2009): Systems chemistry on early Earth. Nature 459, p. 171÷172.
304. Tadeusiewicz R. (1993): Sieci neuronowe. Akademicka Oficyna Wydawnicza RM, Warszawa, Poland.
305. Tadeusiewicz R. (2007): Odkrywanie właściwości sieci neuronowych przy użyciu programów w języku C#. Polska Akademia Umiejętności, Kraków.
306. Tadeusiewicz R. (2009): Neural network as a tool for medical signals filtering, diagnosis aid, therapy assistance and forecasting improving. In Dössel O., Schlegel W. C. (eds.) IFMBE Proceedings, Vol. IV: Image processing, biosignals processing, modelling and simulation, biomechanics. Springer Verlag, vol. 25, Berlin, Heidelberg, New York, p. 1532÷1534.
307. Tajima F. (1983): Evolutionary relationship of DNA sequences in finite populations. Genetics 105, p. 437÷460.
308. Tajima F. (1989): Statistical methods to test for nucleotide mutation hypothesis by DNA polymorphism. Genetics 123, p. 585÷595.
309. Tattersall I. (2003a): Once we were not alone. Scientific American, Special edition: New look at human evolution, p. 20÷27.

310. Tattersall I. (2003b): Out of Africa again … and again. Scientific American, Special edition: New look at human evolution, p. 38÷45.
311. Tavare S., Marshall Ch., Will O., Soligo O., Martin R. D. (2002): Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416, p. 726÷729.
312. Tebelskis J. (1995): Speech Recognition using Neural Networks. Thesis for a degree of Doctor of Philosophy in Computer Science, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.
313. Teraoka S. N., Telatar M., Becker-Catania S., Liang T., Onengut S., et al. (1999): Splicing defects in the ataxia-telangiectasia gene, ATM: underlying mutations and consequences. Am. J. Hum. Genet. 64, p. 1617÷1631.
314. Thomas P. D., Kejariwal A. (2004): Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: Evolutionary evidence for differences in molecular effects. Proc. Nat. Acad. Sci. 101, p. 15398÷15403.
315. Thompson R., Pritchard J., Shen P., Oefner P., Feldman M. (2000): Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. USA 97, p. 7360÷7365.
316. Thorne A., Wolpoff M. H. (1992): The multiregional evolution of humans. Scientific American 266, p. 76÷83.
317. Toomajian C., Kreitman M. (2002): Sequence Variation and Haplotype Structure at the Human HFE Locus. Genetics 161, p. 1609÷1623.
318. Trevors J. T. (1999): Why on Earth: Self-assembly of the first bacterial cell to abundant
319. and diverse bacterial species. World Journal of Microbiology and Biotechnology 15(3), p. 297÷304.
320. Trikka D., Fang Z., Renwick A., Jones S. H., Chakraborty R., et al. (2002): Complex SNP-based haplotypes in three human helicases: implications for cancer association studies. Genome Res 12, p. 627÷639.
321. Turing A. M. (1936): On computable numbers, with an application to the Entscheidungsproblem. Proc. London Math. Soc. Ser. 2, 42, p. 115÷154.
322. Turing A. M. (1938): Correction to: On computable numbers, with an application to the Entscheidungsproblem. Proc. London Math. Soc. Ser. 2, 43, p. 544÷546.
323. Turing A. M. (1950): Computing Machinery and Intelligence. Mind 59(236), p. 433÷460.
324. Twomey J. M., Smith A. E. (1998): Bias and variance of validation methods for function approximation neural networks under conditions of sparse data. IEEE Trans Sys., Man., and Cyber. 28(3), p. 417÷430.
325. Uziel T., Savitsky K., Platzer M., Ziv Y., Helbitz T., et al. (1996): Genomic organization of the ATM gene. Genomics 33, p. 317÷320.

326. Vigilant L., Stoneking M., Harpending H., Hawkes K., Wilson A. C. (1991): African populations and the evolution of human mitochondrial DNA. Science 253, p. 1503÷1507.
327. Von Kobbe C., Karmakar P., Dawut L., Opresko P., Zeng X., et al. (2002): Colocalization, physical, and functional interaction between Werner and Bloom syndrome proteins. J. Biol. Chem. 277, p. 22035÷22044.
328. Von Neumann J. (1951): The general and logical theory of automata. Lecture given in 1948. In Jeffress L. (Ed.) Cerebral Mechanisms in Behavior – The Hixon Symposium, John Wiley, New York, p. 1÷41.
329. Vowles E. J., Amos W. (2006)Ł Quantifying Ascertainment Bias and Species-Specific Length Differences in Human and Chimpanzee Microsatellites Using Genome Sequences. Mol. Biol. Evol. 23(3), p. 598÷607.
330. Wall J. D. (1999): Recombination and the power of statistical tests of neutrality. Genet. Res. 74, p. 65÷79.
331. Wall J. D., Lohmueller K. E., Plagnol V. (2009): Detecting ancient admixture and estimating demographic parameters in multiple human populations. Mol. Biol. Evol. 26, p. 1823÷1823.
332. Wang L., Ogburn C. E., Ware C. B., Ladiges W. C., Youssoufian H., et al. (2000): Cellular Werner phenotypes in mice expressing a putative dominant-negative human WRN gene. Genetics 154, p. 357÷362.
333. Wang W., Seki M., Narita Y., Nakagawa T., Yoshimura A., et al. (2003): Functional relation among RecQL family helicases RecQL1, RecQL5, and BLM in cell growth and sister chromatid exchange formation. Mol. Cell Biol. 23(10), p. 3527÷3535.
334. Weaver S., Baird L., PolyCarpou M. M. (1998): An Analytical Framework for Local Feedforward Networks. IEEE Transactions on Neural Networks 9(3), p. 473÷482.
335. Węgrzyn S., Klamka J. (2000): Kwantowe systemy informatyki. ZN Pol. Śl. Studia Informatica Vol. 21, No. 1 (39), Gliwice.
336. Węgrzyn S. (2010): Molekularne systemy informatyki. Studia Informatica Vol. 31, No. 1, p. 43÷53.
337. Whitley D. (1997a): Evolutionary Computation Models – Representations – Permutations. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C1.4:1÷C1.4:8.
338. Whitley D. (1997b): Evolutionary Computation Models – Search Operators – Mutation – Permutations. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C3.2:5÷C3.2:8.

339. Whitley D. (1997c): Evolutionary Computation Models – Search Operators – Recombination – Permutations. In Bäck T., Fogel D. B., Michalewicz Z. (eds.) Handbook of Evolutionary Computation, Oxford University Press, New York – Oxford, p. C3.3:14÷C3.3:20.
340. Widrow B., Hoff M.E. (1960): Adaptive switching circuits. IRE WESCON Convention Record, New York, p. 96÷104.
341. Wieczorkowski R., Zieliński R. (1997): Komputerowe generatory liczb losowych. WNT, Warszawa.
342. Wilson A. C., Cann R. L. (1992): The recent African genesis of humans. Scientific American 266(4), p. 68÷73.
343. Wolpoff M. H. (1999): Paleoanthropology. Boston, McGraw-Hill.
344. Wooding S., Rogers A. (2000): A Pleistocence population X-plosion?. Human Biology 72, p. 693÷695.
345. Wooding S., Rogers A. (2002): The matrix coalescence and an Application to Human Single-Nucleotide Polymorphisms. Genetics 161, p. 1641÷1650.
346. Wooding S. P., Watkins W. S., Bamshad M. J., Dunn D. M., Weiss R. B., Jorde L. B. (2002): DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 Gene: Implications for Human Population History and Natural Selection. Am. J. Hum. Genet. 71, p. 528÷542.
347. Wu L., Hickson I. D. (2003): The Bloom's syndrome helicase suppresses crossing over during homologous recombination. Nature 426, p. 870÷874.
348. Yamagata K., Kato J., Shimamoto A., Goto M., Furuichi Y., Ikeda H. (1998): Bloom's and Werner's syndrome genes suppress hyperrecombination in yeast sgs1 mutant: implication for genomic instability in human diseases. Proc. Natl. Acad. Sci. USA 95, p. 8733÷8738.
349. Yamaguchi M., Yamamoto K., Miki T., Mizutani S., Miura O. (2003): T-cell prolymphcytic leukemia with der(100)t(1;11)(q21;q23) and ATM deficiency. Cancer Genet. Cytogenet. 146(1), p. 22÷26.
350. Yu C.-E., Oshima J., Wijsman E. M., Nakura J., Miki T., Piussan C., et al. (1997): Werner's Syndrome Collaborative Group : Mutations in the consensus helicase domains of the Werner syndrome gene. Am. J. Hum. Genet. 60, p. 330÷341.
351. Yu N., Zhao Z., Fu Y., Sambuughin N., Ramsay M., Jenkins T., Leskinen E., Patthy L., Jorde L., Kuromori T., Li W. (2001): Global patterns of human DNA sequence variation in a 10-kb region on chromosome 1. Mol. Biol. Evol. 18, p. 214÷222.
352. Yusa K., Horie K., Kondoh K. G., Kouno M., Maeda Y., et al. (2004): Genome-wide phenotype analysis in ES cells by regulated disruption of Bloom's syndrome gene. Nature 429, p. 896÷899.
353. Zadeh L. (1965): Fuzzy sets. Information and Control 8(3), p. 338÷353.

354. Zaher H. S., Unrau P. J. (2007): Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. RNA 13, p. 1017÷1026.
355. Zhang J. (2003): Evolution of the Human ASPM Gene, a Major Determinant of Brain Size. Genetics 165, p. 2063÷2070.
356. Zhivotovsky L. A., Feldman M. W., Grishechkin S. A. (1997): Biased mutations and microsatellite variation. Mol. Biol. Evol. 14(9), p. 926÷933.
357. Ziarko W. (1993): Variable precision rough sets model. Journal of Computer and Systems Sciences 46(1), p. 39÷59.
358. Zietkiewicz E., Votova V., Jarnik M., Koran-Laskowska M., Kidd K., Modiano D., Scozzari R., Stoneking M., Tishkoff S., Batzer M., Labuda D. (1998): Genetic structure of the ancestral population of modern humans. J. Mol. Evol. 47(2), p. 146÷155.
359. Żurada J. (1992): Introduction to artificial neural systems. West Publishing Company, USA.

LIST OF FIGURES

LIST OF TABLES

ARTIFICIAL INTELLIGENCE, BRANCHING PROCESSES AND COALESCENT METHODS IN EVOLUTION OF HUMANS AND EARLY LIFE

Keywords: artificial intelligence, machine learning, computer simulations, branching processes, population genetics, human evolutionary genetics, origins of life

Abstract

The book is composed of two parts, which are preceded by the introduction given in Chapter 1. The introduction presents the genesis of problems considered in the monograph, as well as its organization and objectives. Based on these objectives the main problems discussed in the dissertation are formulated.

Part I, which is as a presentation of methodological apparatus used in the research studies performed by the author, consists of two chapters, Chapter 2 concerning artificial intelligence, and Chapter 3 related to population genetics. Part II shows how the methods described in Part I are applied in author's evolutionary genetics studies. These studies are roughly focused in three areas, the neutral theory of evolution described in Chapter 4, the evolution of humans discussed in Chapter 5, and the origin of life, considered in Chapter 6. The more specific description of particular chapters is given below.

The organization of Chapter 2 is motivated by the natural discrimination between the methods which are inspired by biology, such as artificial neural networks and evolutionary computation, and methods based on formal logic, such as rule-based information systems. It is author's full responsibility that out of many currently studied machine learning methods, he has subjectively chosen in his research neural and evolving systems as those which had arisen from contemplation of life and the rough set theory as the formal logic-based method. However, after this choice has been done and reflected in his studies, the composition of Chapter 2 could not be different. That is also an explanation why the last section in this chapter is a case study – its goal is to illustrate how in one practical application, all these three approaches have found their place.

Chapter 3 is a brief presentation of population genetics models, which are used, in addition to machine learning and computer simulations, in author's studies considered in Part II. Comparing the content of this chapter with what is classically understood as a population genetics, the reader will notice that except typical material, such as the Wright-Fisher model of a genetic drift, drift-mutation-selection interplay, and the coalescent method, the chapter also contains a section about genealogy of branching processes. This latter is again the subjective choice, which has been made before writing of the book was started. It was made at the time when the author, inspired by an excellent Kimmel's and Axelrod's book, has introduced to population genetics-related research the branching processes models, in particular the O'Connell model of branching processes genealogy.

Chapter 4, entitled "Theory of Neutral Evolution", after presenting introductory material concerning Kimura's theory of neutral molecular evolution and its relation to the Darwinian selection-driven evolution, focuses on how this theory can be used in search for signatures of natural selection at molecular level. The neutrality tests, which have been designed for detection of such selection are presented, before the case study on that issue is given. The problems with interpretation of the results are the starting point for development of two author's methods: multi-null-hypotheses method and the machine learning-based quasi dominant rough set approach.

In Chapter 5, the human evolution is the central point. Within this field many approaches are used for inferring the past of our species, including paleontology and evolutionary genetics. On the background of two competing theories of modern human origin, the multiregional and the recent out-of-Africa hypotheses, there are presented studies concerning detection of past population expansion using classical and author's neural network-based tests. This material is followed by reporting on the research concerning the mitochondrial DNA record. In particular, it is shown how the date of the root of mitochondrial DNA polymorphism is estimated using the O'Connell and the Wright-Fisher models in forward-time computer simulations of slightly supercritical branching processes. Additionally, in Chapter 5 it is demonstrated how the criticality of branching processes has been used for modeling the decay of hypothetical admixture of Neanderthal mitochondrial DNA in a gene pool of the Upper Paleolithic anatomically modern humans. This issue is currently hot debated in the light of results from the Neandertal Genome Project and discussions about interbreeding between *H. sapiens* and *H. neanderthalensis*.

As Chapter 5 was focused on evolution, which took place less than million years ago, the Chapter 6 speculates about the times almost as ancient as the age of Earth. The point of gravity of Chapter 6 is computer science contribution to the problem of how life has emerged. With that regard, three models are discussed. The first is the Demetrius-Kimmel complexity threshold model supplemented by the author to include hydrolysis of RNA strands caused by

phosphodiester bond break reaction. The second is the modification of the Niesert compartment model with random segregation of genetic material. The third is the Monte-Carlo model proposed by Ma and collaborators in 2007, and supplemented by simulation of non-enzymatic template-based RNA recombination process, which seemed to be significant in the emergence of the RNA World.

Finally, these three application-oriented chapters which constitute Part II of the book, are followed by Chapter 7, which gives the opportunity, not only to summarize the issues discussed in the whole monograph, but also to go beyond that material, by speculating on philosophical matters, which naturally occur when the artificial intelligence is considered.

METODY SZTUCZNEJ INTELIGENCJI, PROCESÓW GAŁĄZKOWYCH I KOALESCENTU W BADANIACH EWOLUCJI CZŁOWIEKA ORAZ WCZESNEGO ŻYCIA

Słowa kluczowe: sztuczna inteligencja, uczenie maszynowe, symulacje komputerowe, procesy gałązkowe, genetyka populacyjna, ewolucyjna genetyka człowieka, początki życia

Streszczenie

Niniejsza monografia składa się z dwóch części, które są poprzedzone wstępem zawartym w Rozdziale 1. We wprowadzeniu przedstawiono genezę problemów rozważanych w monografii, jak również jej organizację oraz cele. W oparciu o te cele zostały sformułowane główne problemy rozprawy.

Część I, będąca prezentacją aparatu metodologicznego wykorzystywanego w badaniach naukowych autora, składa się z dwóch rozdziałów: Rozdziału 2 dotyczącego sztucznej inteligencji i Rozdziału 3 na temat genetyki populacyjnej. W Części II pokazano jak metody opisane w Części I są wykorzystywane w badaniach prowadzonych przez autora w zakresie genetyki ewolucyjnej. Badania te skupiają się wokół trzech dziedzin: teorii neutralnej ewolucji, opisanej w Rozdziale 4, ewolucji człowieka, opisanej w Rozdziale 5, oraz pochodzenia życia, rozważanego w Rozdziale 6. Bardziej szczegółowy opis poszczególnych rozdziałów znajduje się poniżej.

Organizacja Rozdziału 2 jest motywowana naturalnym zróżnicowaniem pomiędzy metodami, które są inspirowane przez biologię, takimi jak sztuczne sieci neuronowe i obliczenia ewolucyjne, oraz metodami opartymi o logikę formalną, takimi jak regałowe systemy informacyjne. Autor bierze pełną odpowiedzialność za to, że spośród wielu aktualnie wykorzystywanych metod uczenia maszynowego, wybrał w swoich badaniach systemy neuronowe i ewolucyjne, jako te które wyrosły z kontemplacji życia, oraz teorię zbiorów przybliżonych jako metodę opartą na logice formalnej. Jednakże, po dokonaniu tego wyboru odzwierciedlonego w jego badaniach, kompozycja Rozdziału 2 nie mogła być już inna.

Wybór ten wyjaśnia również dlaczego ostatnia sekcja tego rozdziału jest poświęcona studium przypadku – jej celem jest zilustrowanie jak wszystkie te trzy podejścia znajdują swoje miejsca w jednym praktycznym zastosowaniu.

Rozdział 3 jest zwartą prezentacją modeli genetyki populacyjnej, które wykorzystywane są, obok uczenia maszynowego i komputerowych symulacji, w badaniach autora rozważanych w Części II. Porównując zawartość tego rozdziału z klasycznie ujmowaną genetyką populacyjną, czytelnik zauważy, że oprócz typowego materiału, takiego jak model dryfu genetycznego Wrighta-Fishera, współdziałania dryfu, mutacji i selekcji, oraz metody koalescentu, rozdział zawiera sekcję na temat genealogii procesów gałązkowych. To ostatnie zagadnienie jest znowu subiektywnym wyborem, dokonanym przed rozpoczęciem pisania książki. Decyzja została podjęta, kiedy autor, zainspirowany przez doskonała książkę Kimmla i Axelroda, wprowadził do swych badań z zakresu genetyki populacyjnej modele procesów gałązkowych, a w szczególności model O'Connella dotyczący genealogii procesów gałązkowych.

Rozdział 4, zatytułowany „Teoria Ewolucji Neutralnej", po przedstawieniu materiału wstępnego dotyczącego teorii Kimury zwanej teorią neutralnej ewolucji molekularnej jak również jej związków z Darwinowską ewolucją napędzaną przez selekcję, rozważa jak teoria ta może być wykorzystana w poszukiwaniu znamion selekcji naturalnej na poziomie molekularnym. Pokazano testy neutralności, które zostały zaprojektowane do wykrywania takiej selekcji, a następnie ich wykorzystanie w studium przypadku. Problemy interpretacji rezultatów tych testów stanowiły punkt wyjścia do rozwinięcia dwóch autorskich metod: metody wielu hipotez zerowych, oraz metody opartej na uczeniu maszynowym z użyciem podejścia quasi-dominujących zbiorów przybliżonych.

W Rozdziale 5 centralnym punktem jest ewolucja człowieka. W tej dziedzinie zaproponowano wiele podejść by odkryć przeszłość naszego gatunku, w tej liczbie, metody paleontologiczne i genetyczne. Na tle dwóch konkurujących teorii pochodzenia człowieka współczesnego, hipotezy wieloregionalnej oraz hipotezy pożegnania z Afryką, zaprezentowane są badania mające na celu wykrycie przeszłych okresów ekspansji populacji w wykorzystaniem metod klasycznych, oraz, opartej o sieci neuronowe, metody autora. Następnie, przedstawiono raport z badań na temat zapisu mitochondrialnego DNA. W szczególności, pokazano jak estymowano epokę korzenia polimorfizmu mitochondrialnego DNA z wykorzystaniem modeli O'Connella oraz Wrighta-Fishera w symulacjach komputerowych lekko nadkrytycznych procesów gałązkowych. Ponadto, w Rozdziale 5 pokazano jak wykorzystać krytyczność procesu gałązkowego do modelowania zaniku hipotetycznej domieszki neandertalskiego mitochondrialnego DNA w puli genów ludzi anatomicznie współczesnych Górnego Paleolitu. Ta kwestia jest aktualnie gorąco

dyskutowana w świetle rezultatów Projektu Neandertalskiego Genomu oraz dyskusji na temat krzyżowania pomiędzy *H. sapiens* i *H. Neanderthalensis.*

O ile Rozdział 5 był poświęcony ewolucji działającej w okresie mniej niż milion lat wstecz, Rozdział 6 spekuluje na temat czasów prawie tak starych jak sama Ziemia. Punktem ciężkości Rozdziału 6 jest wkład informatyki do problemu powstania życia. W tym kontekście dyskutowane są trzy modele. Pierwszy, to model granicy złożoności Demetriusa-Kimmla, uzupełniony przez autora tak, by uwzględniał hydrolizę łańcuchów RNA spowodowaną przez reakcję rozpadu wiązania fosfodiestrowego. Drugi, to modyfikacja kompartmentowego modelu Niesert z losową segregacją materiału genetycznego. Trzeci, to model Monte-Carlo zaproponowany przez Ma i współpradcowników w 2007, i uzupełniony przez symulacje procesu nieenzymatycznej opartej o wzorzec rekombinacji RNA, który to proces wydaje się być znaczący w powstaniu świata RNA.

Na koniec, po tych trzech zorientowanych na zastosowania rozdziałach, które stanowią Część II monografii, Rozdział 7 stanowi okazję nie tylko do podsumowania problemów poruszanych w całej rozprawie, ale również do wyjścia poza ten materiał, poprzez rozważanie kwestii filozoficznych, które naturalnie się pojawiają w myśleniu o sztucznej inteligencji.