Algorithms for modeling of the evolution of complex stochastic genetic systems.

Tomasz Wojdyła

Institute of Informatics

Silesian University of Technology

Supervisor: **prof. dr hab. inż. Marek Kimmel**

A thesis submitted for the degree of

PhilosophiæDoctor (PhD)

2011 October

To My Parents

Contents

1

Introduction

In the late 1850s Charles Darwin and Alfred Wallace (32, 33) introduced a new theory of evolution based on the natural selection and changed the human perspective on the origin of life. A few years later Gregor Mendel formulated his Laws of Inheritance giving a foundation of genetics (127). Philosophical and religious controversies of the newly announced theories were not conducive to rapid progress of studies. Only in the beginning of the 20th century scientists "rediscovered" this discipline of science. We can notice increased scientific activity to explain mechanisms behind the evolution taking place in that time. These studies gave a birth of population genetics. In the beginning, the population genetics focused on the statistical aspects of the evolution forces leading to very simplified, but important, theorems (i.e., the Hardy-Weinberg principle (31)) and models (i.e. the Wright-Fisher model (52, 187)). Discovering the coalescent theory (90, 110, 171) in the 1980s along with popularization of computer calculations were another milestone in genetic studies leading to many Monte-Carlo simulation (128, 129) models and methods. Constant development of computers and biology, especially DNA sequencing methods, caused that currently more and more complex models, including those operating on real data, can be applied. Currently, medical significance of the possible results of population genetics projects determine enormous financial and intellectual support. At least a few of these recent projects deserve to be mentioned due to their contribution in our knowledge of the human genetics and huge resources spent on them. The Human Genome Project (1, 95, 96) gave us a database with more than 20000 identified human genes. The International HapMap Project (2, 25) goal is to identify and catalog entire haplotype map of the

human genome. The goal of The 1000 Genome Project (3, 26), started in 2008, is to develop a deep human genetic variation catalogue. All of three mentioned above projects deliver us open-access real data that we can use in our analysis.

As we see, despite a huge progress that has been made over last several decades, population genetics is still a very young discipline of science. Some aspects of the main evolution forces (such as natural selection, mutation or recombination) are already known very well but details of their interaction are unclear. Unfortunately, any realistic, and for this reason very complex, models tend to be too complicated for mathematical and statistical analysis even with use of computing power.

In this dissertation the attention is focused on the complex analytical stochastic systems refined by specialized computer algorithms. We present three such models investigating some aspects of interactions between recombination, mutation, genetic drift and population growth. The models are used to prove the following **dissertation theses**:

- It is possible, using a non-simulation approach applied to the mathematical Moran model, to answer the question of the recombination identifiability, at least in the means of the relationships limited to a set of distributions, which jointly characterize allelic states at a number of different loci.

- It is possible, using a recursive algorithm, to calculate the exact distribution of the time to the MRCA of a large sample from a population evolved under any growth scenario with the time efficiency of the method allowing for analysis of large human populations.

- It is possible to build a non-simulation model of demographic interactions between many populations or species that can, in some applications, replace the simulation-based approach.

The first model describes the asymptotic behavior of a well-known Moran model (44) with mutation, genetic drift and recombination along multiple loci. The second model investigates the distribution of the time to the most recent common ancestor (MRCA) of a sample in the Wright-Fisher model with variable population size. Finally, the last model is a demographic network of populations with time-varying sizes experiencing merges, splits and migration events. All of these models are theoretical models too

complicated for common statistical or mathematical analysis. We deliver specialized computer algorithms that allow us to explain interesting properties of these models and obtain valuable results.

The dissertation begins, in Chapter 2, with description of the basic evolutionary mechanisms along with methods of modeling them. In this chapter we also present usually used approaches for modeling of the evolution of genetic systems.

Chapter 3 contains the formulation of the aim and the theses of this dissertation along with the explanation of the reasoning behind each investigated model.

Chapters 4, 5 and 6, containing the main contribution of the dissertation, present all studied models. Each of these chapters is dedicated to a specific model and contains detailed description of the model along with all methods and computer programs used to examine these models and obtained results. Chapter 4 describes the mathematical preliminaries and asymptotic behavior of the refined Moran model with mutation, genetic drift and recombination along multiple loci. We deliver a computer program that allows us to analyze the model for non-trivial cases (with number of loci greater than two). We also present obtained results, including spectral gap analysis and numerical studies of the asymptotic behavior, and compare this model with the Hudson's Wright-Fisher coalescent model (90, 91). In Chapter 5 we introduce a new method, which is a computer algorithm based on the dynamic programming (15, 29), to calculate the exact distribution of the time to the MRCA for a sample drawn from a population with time-varying size. We use our method on different data sets including real human populations, artificially generated populations with bottleneck events and populations evolved according to the branching process. In Chapter 6 we present a computer program realizing a complex demographic network model. The model takes into account the most important demographic events, such as splits, merges and migrations. We do not assume any specific mutation model. We allow for mutation rates and population sizes change over time. We also discuss several model extensions useful for several commonly-used mutation models and show the example applications of the model.

In the last chapter of the dissertation, Chapter 7, we sum up and discuss results of the thesis.

2

Preliminaries

2.1 Main evolutionary mechanisms and their models

In this section we discuss the basics of several evolutionary forces along with methods used to model them. We do not describe all known evolutionary mechanisms but we concentrate on the most influential instead. We do not know many aspects of these omitted mechanisms yet. Moreover, the interactions between discussed forces seem to be very complex and usually extremely hard to model efficiently. Therefore, almost all existing models skip all non-listed in this paragraph forces.

2.1.1 Mutation

Mutations are considered as one of the most important genetic force. The reasoning of this statement is simple, mutations are the main force that adds genetic diversity to the gene pool of a population. The strength of mutation can be measured by the mutation rate parameter, denoted by μ, which is the probability that a mutation event occurs in a single individual (or gene) in one generation. We also assume that $\theta = 4N_e\mu$ is the population mutation rate, where N_e is the effective population size. The average human mutation rate was estimated to be equal to about $2.5 \cdot 10^{-8}$ mutations per nucleotide site (135). Detailed studies showed that the value of mutation rate changes between different regions and there are regions with extremely high mutation rate, i.e. the mutation rate of human mitochondrial DNA is estimated, depending on the used method, to be equal to between $3 \cdot 10^{-6}$ (99) and $2.7 \cdot 10^{-5}$ (89, 148). Moreover, the mutation rate can significantly vary even in the range of the same region, especially in

the present of recombination hotspots (81, 88). Unfortunately, the interactions between mutation and recombination are still very unclear and we do not know any method to estimate such region-varying mutation rate. Thus, we usually assume that the mutation rate in a model is constant and depends on the modeled region. Such average mutation rate can be estimated by using two well-known estimators, Watterson's estimator (181) or Taijma's estimator (172).

Two main models of mutation are usually considered. The most commonly used model is the Single-nucleotide polymorphism (SNP) model where all possible mutation positions (and sometimes also their variations) in a DNA fragment are described. A single mutation event in the model is a point mutation that replaces one nucleotide with another. We can distinguish many variants of this model. The most general division contains an infinite-site model, a finite-site model and an infinite-allele model. The second kind of models is the microsatellite model based on repeats of a short DNA sequence called tandem repeats.

A mutation model applied to the model of population determines the allelic space of individuals. Throughout the dissertation we denote the countable set of allele types as $\mathbb{A}$ and the number of allele types as $N_{\mathbb{A}}$.

Infinite-site model

Mutation in this model can occur at any site of a long DNA sequence, but only once per site. Hence, a new mutation always determines a new variable site (SNP) and two different individuals can have the same mutation only via inheritance. Usually we assume that each site can take only two possible alleles and we describe an individual as a sequence of 0s and 1s (where, i.e., 1 stands for an occurrence of mutation at the site and 0 otherwise). Sometimes we add the exact position of the SNP, usually as a number from the [0,1] range. There are no recurrent mutations allowed and the number of mutations is equal to the number of variable sites. From the biological point of view the model is justified by an assumption that the probability of mutation is very low and it is unlikely that two or more mutations occur at the same position.

Finite-site model

A finite-site model is the most popular model of mutation. In this model we know the exact number and positions of sites and we assume that a mutation can occur at any

site. We usually allow for recurrent mutations and we almost always do not assume any difference between sites. The number of sites may differ from few (even one) to a very large number (for, i.e., a long haplotype DNA sequence). A finite-site model with many enough sites is a good and commonly-used approximation of the infinite-site model. The model allows for any finite number of alleles per site but this number is usually equal to two. A very important Kimura K80 model (109) adapts a finite-site model with four possible allele states per site and different transition probabilities.

Infinite-allele model

In this model we assume that an occurrence of mutation in any individual creates a completely new, not observed before, allele. Thus, all we know from the model is which individuals have identical alleles. This model is biologically inspired by isozymes, which are differently charged forms of an enzyme (80).

Microsatellite model

Microsatellite regions consist short tandem repeats. The number of repeats varies from 10 to 100 and each repeat contains 1-6 DNA base pairs. Microsatellites were the first genetic markers ever used and they are still very popular. The main reasons for that are wide existence of microsatellite regions across the genome (especially CA repeats) and a large polymorphism of allele lengths at a microsatellite loci. Thus, microsatellites are commonly used to determine paternity, solve pedigree linkage problems, identify individuals or even as neutral mendelian markers (98).

Constructing of the population genetics model with tandem repeats requires a microsatellite mutation and an evolution model to be established (40). The most widely used model is the stepwise mutation model (SMM). The model assumes that a single mutation either adds or removes exactly one repeat from the current length of the microsatellite allele (144). The SMM model assumes the same probability for a single forward or backward step in the allele length omitting any other molecular dynamics factors having impact on the mutation. This simplicity causes that we cannot obtain the stationary distribution of allele sizes from the SMM model. Therefore, many extensions to the basic SMM model have been added. These refined models, named as a general stepwise mutation models (GSSM), can be obtained, i.e., by: allowance of multiple steps in a single mutation event (106, 138), introduction of the upper limit

of the length of allele (47, 137), introduction of asymmetric probabilities for possible directions of a single step (106, 188).

2.1.2 Genetic drift

Genetic drift is a mechanism that stays behind changes of the allele frequencies in a population caused by mating of individuals from that population (coalescence events). In the early models the effect of the drift was discarded by assuming an infinite size of population. In that case, and under a few other assumptions, the allele frequencies are in equilibrium and depend on the genotype frequencies (the Hardy-Weinberg principle (31)). If we assume a more realistic model, where the population size is finite, we can notice that mating changes the allele frequencies in two ways. Mixing of the genotypes of two individuals allows to spread some alleles in the gene pool. On the other hand, coalescence causes the extinction of some number of lineages (and genes) leading to an average decrease of the heterozygosity by $\frac{1}{2N_e}$ per generation (161). As we see, the effect of genetic drift depends on the size of population and can be very strong for small populations. If the mentioned heterozygosity decrease is balanced by the effect of mutations, we say that the population is in mutation-drift equilibrium.

We need to introduce a sampling scheme to the model in order to model genetic drift. This scheme is usually uniformly random (the Wright-Fisher model, the Moran model). A non-random sampling scheme can cause effects similar to the ones created by a non-neutral selection. Thus, one should use neutral mutation models only to analyze the genetic drift.

2.1.3 Recombination

Recombination is yet another very important evolutionary mechanism. We distinguish two main types of the recombination events: crossover and gene conversion. Crossover is an exchange of a DNA fragment between paired chromosomes inherited from each of one's parents. These events may be very useful for geneticists to track the linkage between genes (genes situated closely to each other should be a part of the same recombining fragment more often than the one's that are far from each other). Gene conversion is a replacement of a DNA fragment of one chromosome by a fragment from the second paired chromosome. Recombination events usually occur at the specific location, named as a recombination hotspot (84).

Unfortunately, our knowledge about recombination is still relatively small. We can point at population genetics effects that are, with high probability, caused by recombination (i.e., the loss of SNPs density after human departure out of Africa (169)) but we cannot explain them yet. Recombination, being a very complex mechanism, is also hard to model and analyze. There exist several methods, mostly the backward-time models of a crossover recombination, that allow us to study recombination. Griffiths (66) and Hudson (90) introduced recombination into the coalescent theory by allowing a coalescence event to be, with the recombination rate r, a crossover recombination. We assume that the recombination rate r is the probability that two individuals experience a recombination event during their coalescence. Based on the mentioned models, a few new backward-time models of recombination have been developed (38, 80, 179). Current forward-time recombination analysis rely on computer simulations (152). In the recent studies geneticists try to develop a method to estimate the recombination rate from the real data. Hudson (92) described a method based on the coalescent theory that uses an approximation of the likelihood surface. McVean improved this method by introducing recurrent mutations (125), by allowing for the variable recombination rate over the DNA region (126) and finally, by adding the recombination hotspot model (7). There are only few developed models that are used to explain mechanisms behind the recombination (i.e., the loss of the SNPs density (164)).

2.1.4 Migration

Migration (called also as gene flow) is a transfer of genetic material between populations. Migration may lead to introducing or reintroducing genes to the population, increasing the genetic variation of that population. By moving genes around, migration can make distant populations genetically similar to each another, hence reducing the chance of speciation. The less gene flow between two populations, the more likely that two populations will evolve into two species.

Although migration is often a continuous process between two populations, we usually model it as a single event when a part of one population merges with the second population. This approach allows to sustain the main effects of the migration without unnecessary complication of the model.

2.1.5 Natural selection

We know that alleles of genes of individuals that are better suited to the environment (by possessing better phenotype features) will increase their frequencies in the gene pool. An evolutionary mechanism that stays behind this process is called as natural selection. The ability of individual to increase its allele frequency can be measured by the fitness parameter, usually denoted as w. In the early studies fitness was identified with the ability to survive. In the more recent years this point of view has changed and the ability to reproduce is considered now to be the main factor of the fitness. The fitness value depends also on the number of individuals in a population and on the frequencies of various alleles (44).

Natural selection is the key of the evolutionary process. The fundamental theorem of natural selection (52) claims that the rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at that time. In fact, the mean fitness increase may not be true if we consider a non-random mating or a multiple loci dependance mating (45). Biologically we interpret natural selection as a form of egoism where individuals prefer a survival of units of selection (kin groups) with the same genes (23, 34, 75, 76).

2.2 Basic evolutionary population models

In this section we shortly describe several models of evolution of a population. All presented models are basic, well-known and commonly used concepts and require significant refinement to be applied in order to model any more sophisticated aspects of the evolution. Each of the described models is a base for models used in our studies. In Chapter 4 we will introduce the Moran model with mutation and recombination and compare some aspects of this model to the Wright-Fisher model. Our genetic drift (Chapter 5) and demographic (Chapter 6) models use the Wright-Fisher approach for modeling of the genetic drift. We will also discuss and present results of applying our algorithm calculating the time to the MRCA to the Galton-Watson branching process (Chapter 5).

2.2.1 Wright-Fisher model

The Wright-Fisher model (52, 179, 187) describes the transmission of genetic material in a population over the generations. The population has constant size and usually consists $2N$ genes corresponding to N diploid or $2N$ haploid individuals. Rarely, a haploid population consists only N individuals. In a diploid population the number of female individuals N_f does not have to be equal to the number of male individuals $N_m = N - N_f$. Generations in the model are discrete and do not overlap. In case of human population we usually assume that one generation lasts 25 years. All individuals from the population in generation i are replaced in generation $i+1$ by their descendants. In the haploid model each individual in daughter generation is chosen by random draw with replacement from the population in mother generation. In the diploid model, each individual from daughter population has two randomly chosen (with replacement) ancestors from mother population (one male and one female) but inherits only one gene, either from mother or from father. The probabilities of inheriting from mother and from father are equal.

In the basic Wright-Fisher model we assume that individuals in a population are not affected by any other evolutionary force. In particular we assume that there is no recombination between genes and genes cannot be changed by mutation. Also all individuals in the population are equally fit to the environment conditions (selection has no effect).

The probability that two individuals in generation $i+1$ have the same ancestor in generation i is equal to $\frac{1}{2N}$. Thus, the average time to the MRCA of two individuals is equal to $2N$ generations. The probability distribution of the number of descendants v in generation $i+1$ of the individual from generation i is given by a binomial distribution:

$$P(v = k) = \binom{2N}{k} \left(\frac{1}{2N}\right)^k \left(1 - \frac{1}{2N}\right)^{2N-k} \tag{2.1}$$

Therefore, the mean number of descendants is equal to 1 with the variance $1 - \frac{1}{2N}$.

2.2.2 Moran model

The Moran model (132, 179) is a model very similar to the Wright-Fisher model. The only difference between these two models lays in the fact that the Moran model allows for overlapping generations with only one coalescent event per generation. In

continuous-time approach of the Moran model each individual has its own lifetime usually determined by a Poisson distribution. The probability that two individuals share the MRCA in the previous generation is equal to $\frac{1}{\binom{2N}{2}} = \frac{1}{N(2N-1)}$. Thus, the time to the MRCA is a geometric distribution with mean equal to $N(2N-1)$.

2.2.3 Branching process

The branching process (6, 78, 105) is a stochastic process that models the growth of a population of particles. We start at time $t = 0$ with a population of size $Z(t) = 1$. In most branching processes, when one of the particles (lets say, ith) from the population dies, the particle is replaced by its progeny. Other processes, so called Jagers-Crump-Mode branching processes (97), allow for production of a new particle during a lifetime of its parent. We will exclude these models from further discussion.

Each new particle begins a new branching process. The number of progeny ζ_i is usually given by a Poisson distribution. Depending on the value of the mean number of progeny $m = E(\zeta_i)$, we define supercritical ($m > 1$), critical ($m = 1$) and subcritical ($m < 1$) branching processes. The probabilities of extinction of subcritical and critical branching process are equal to 1 and are lesser than 1 in the case of supercritical process. The expected asymptotic values of the population size for supercritical, critical and subcritical branching processes are ∞, 1 and 0, respectively.

In the most general case, the branching process is time-continuous and age-dependent. It means that each particle has its own time of life given by the lifetime distribution function $G(t)$. Such process is called as a Bellman-Harris branching process (16). The Bellman-Harris process is, with two exceptions, a non-Markovian process. The first exception occurs when the $G(t)$ function is exponential and the second exception is a Galton-Watson branching process (51). The Galton-Watson process is the most commonly used branching process with discrete states and discrete generations with constant lifetime of all particles.

Using branching process is a simple method to model parent death - progeny birth dynamic. Thus, branching processes found a wide application in modeling populations of biological cells, genes or biomolecules. Branching processes were firstly used by Watson and Galton to estimate the probability that a human family extinct (180). Other well known results obtained by using branching process are the derivation of a formula for the probability of fixation of a new advantageous mutation (73) or applying

O'Connell model (141) to the estimation of the time to when the female ancestor of modern humans (mitochondrial Eve) lived. For more example see (105).

2.3 Classic methods of modeling of genetic system's evolutionary dynamic

In this section we describe the most important methods used to build the genetic evolutionary systems. Each of these methods is a basic foundation that needs to be refined before applying in order to receive a complete solution of a specific problem. We begin with dividing of all genetic systems into two groups with respect to the chosen approach to the time change in the model. Backward-time and forward-time methods significantly differ from each other and it is important to understand how these differences affect the model. Next, we present the coalescent theory – the main backward-time method of modeling. We will use this approach to compare our Moran-based model to the standard Hudson's Wright-Fisher model. The coalescent theory stays also behind the derivation of the population dynamic model in our demographic network. Models based on a Markov process are very common in the population genetics. We will discuss several aspects of these methods, especially that the main model described in this dissertation widely uses the theory of Markov chains. Monte Carlo methods are very often used to obtain results from computer simulations. We will use this approach several times in this dissertation (i.e., to compare our algorithm calculating the time to the MRCA with simulation methods). Finally, the diffusion approximation is an important mathematical concept that requires to be shortly explained in this thesis, although we will not use it explicitly in any of our models.

2.3.1 Difference between backward-time and forward-time methods

When we construct a model of the evolution of genetic system we need to decide which approach guarantees us obtaining desired results in the most efficient way. If modeled population changes over time, we can model it either by backward-time or forward-time method.

Backward-time approach focuses not on a whole population, but on a sample chosen from this population, usually at the present time. Applying backward-time model requires two steps to be executed. In the first step we build the coalescent tree to the

MRCA of the sample. To obtain the tree we use a stochastic process characterized by evolutionary forces that we take into consideration in the model. In the second step we begin from the found MRCA of the sample and apply forward-time process that assigns genetic information to individuals in the tree.

Forward-time approach in turn is a completely population-based method. It starts with the initialization of the population and follows with generation-to-generation evolution to the final generation (usually being the present time). In most cases the sample is chosen from the last generation.

Forward-time methods are more intuitive and allow to model evolutionary mechanisms in a much simpler way. Unfortunately, they require a whole population to be managed. Thus, computer simulations based on these methods may be extremely time and memory inefficient. Backward-time approaches omit this problem with the cost of computational complexity. Most of the evolutionary forces have already been applied to the coalescent backward-time models (i.e., recombination (66, 90) or simple selection (27, 114)). Although, more complex problems (like realistic human diseases (151, 162)) can only be modeled by the forward-time methods. Finally, the backward-time methods are often based on approximations and equilibrium assumptions and are supposed to work only for certain parameter ranges (178), such as low recombination and mutation rates. The limits of the backward-time approach and constant increase of the computer powers justify development of the forward-time methods (simuPop (152), EASYPOP (11), TreesimJ (143)).

2.3.2 Coalescent-based methods

Development of the coalescent theory (66, 90, 110) revolutionized the methodology of modeling of the population evolution. Before that time the only available approach performed the generation-to-generation whole-population forward-time calculations. As we mentioned earlier, the coalescent theory focuses on a sample constructing the coalescent tree of this sample to its MRCA. Thus, we narrow calculations down to only few generations and individuals that contain any genetic information from the individuals from the sample. Most coalescent approaches adapts the Wright-Fisher model. Based on the knowledge of the population size, the expected life length and mating scheme, we can estimate the mean and the distribution of the time (counting backward from the

present) to the first coalescent event in the sample. For example for standard Wright-Fisher model and the sample of size n ($n \ll 2N$) this time is equal to $\frac{4N}{n(n-1)}$ leading to the following formula for the expected time to the MRCA (T_{MRCA}):

$$E(T_{MRCA}) = \sum_{i=2}^{n} \frac{4N}{i(i-1)} \tag{2.2}$$

Hence, obtaining of the genealogy tree of the sample (and introducing genetic drift to the model) is straightforward.

The most common method of adding mutation to the model is by calculating of the total length of branches of the obtained genealogy tree. The mutation rate μ usually describes a chance for mutation appearance in one individual in one generation time, what is equivalent to the total length of branches equal to 1 (generation). Therefore, estimating the total number of mutation events in the history of sample is easy by applying a random (usually Poisson) distribution with a given parameter. Then, all mutation events are distributed over the tree branches according to the lengths of these branches (the length of the branch is proportional to the probability that the mutation event will occur on this branch). The value of the expected total length of branches TBL is twice the sum of the average waiting times for each coalescent event (74):

$$E(TBL) = \sum_{i=1}^{n-1} \frac{4N}{i} \tag{2.3}$$

We can also introduce other evolutionary mechanisms to the coalescent model by refining the first step of the model (tree building). Hudson (90) described a method of introducing recombination to the model. We can distinguish a second kind of events in the sample – recombination events between the individual from the sample and other individual (sometimes we assume that $n \ll 2N$ and the second recombining individual is chosen from outside of the sample). These events lead to the shortening of the average time to any event in the sample. When the event occurs, we can decide (based on the value of the recombination rate and the sizes of the population and of the sample) which kind of event we deal with and act according to that. In order to model recombination we usually distinguish more than one locus in each individual and the recombination event leads to the separation of the genealogical lineages of different loci from the same individual (chosen from the current generation).

The coalescent theory allowed to develop likelihood methods based on the estimation of the probability of occurrence of a given genealogy tree G under a given population model. For example, in the standard Wright-Fisher model

$$P(G|N) = \prod_{i=2}^{n} \exp\left(-u_i \frac{1(i-1)}{4N}\right) \frac{1}{2N}, \tag{2.4}$$

where u_i is the time interval between two consecutive coalescent events $(i-1)$th and ith (counted from the MRCA) (110). Likelihood methods allow us to estimate the values of the different model parameters even in very complicated models (50). The likelihood $L(a)$ of the parameter a is given by the formula (36):

$$L(a) = \sum_{G} P(\text{Data}|G)P(G|a) \tag{2.5}$$

We usually use Bayesian (12, 115) or Monte Carlo (125) methods for likelihood analysis.

2.3.3 Markovian stochastic processes

A Markovian process (122) is a stochastic process that satisfies the Markov property. Informally it means that the future probabilities of the process are determined only by its most recent values. Assume that at time t a stochastic process X can take, with some probability, a value $X_t = x(t)$. We denote the times for which the Markov process is defined as $t_0 < t_1 < ... < t_{n-1} < t_n$, where t_{n-1} is the present time and t_n is the future time. Satisfying of the Markov property means that the following formula holds (147):

$$P(X_{t_n} = x(t_n)|x(t_{n-1}), x(t_{n-2}), ..., x(t_0)) = P\left(X_{t_n} = x(t_n)|x(t_{n-1})\right) \tag{2.6}$$

Markov chain is a Markov process with a finite number of states. In the dissertation we distinguish discrete-time Markov chains (if the chain is defined for a discrete set of times) and continuous-time Markov chains (otherwise).

Let the transition probability matrix $P(t) = \{p_{ij}(t)\}$ be a square stochastic matrix with the elements being the probabilities that X moves from the state i to the state j in the time interval equal to t: $p_{ij}(t) = P(X_t = j|X_0 = i)$. Based on the Chapman-Kolmogorov equation (185), we get

$$P(X_{t_n} = x(t_n)|x(t_s)) = \int_{-\infty}^{\infty} P(X_{t_n} = x(t_n)|x(t_r))P(X_{t_r} = x(t_r)|x(t_s))\mathrm{d}x(t_r), \tag{2.7}$$

where $n > r > s$. Thus, $P(t+s) = P(t)P(s)$.

As we can notice, the Markovian process (and the mathematical theory that stays behind it, especially the analysis of the ergodicity) is a perfect method to model the evolution of the population. In that case the population size is considered as a Markov process state and the removal or addition of the individual as a change of the state. The discrete-time process is usually applied if the time between the change of the state models a single generation. Described process is called as the population process (133). Most of the branching processes are population processes (105).

Markov process is not limited to modeling the population growth only, i.e., we can use Markov process to model mutation (SMM microsatellite model (144)). In this dissertation we use Markov chain to describe the dynamic of the distribution of individuals in population modeled by the Moran model with mutation, recombination and drift.

2.3.4 Monte Carlo methods, MCMC method

Monte Carlo method (128, 129) is a mathematical paradigm that allows us to obtain accurate results from the model even if the exact calculations cannot be applied due to efficiency or complexity issues. The method was firstly used by Ulam in 1946. The basic idea of the method relies on averaging of the results obtained from sample, randomly generated single realizations of the model. The accuracy of the method strongly depends on the number of averaged experiments and is of the order of $\sqrt{k}$, where k is the number of experiments (183). Monte Carlo methods are broadly used in many computational simulations including computational biology (131), applied statistics (166) or genetics (68). We will use this method to calculate the distribution of the time to the MRCA in the Galton-Watson branching process (Chapter 5).

One of the most important application of this method to the population genetics is the Markov Chain Monte Carlo method (MCMC) (58, 77). The MCMC method uses the theory of ergodicity of Markov chains. Assume that Ω is a complex finite probability space with the stationary distribution π. The MCMC method builds (simulates) an irreducible non-periodic Markov chain with a given state space $S = \Omega$, a simple transition matrix P and a stationary distribution π. As a result of the MCMC method realization we obtain a set of fair samples. The basic idea (known as Metropolis

MCMC) was introduced by Nicholas Metropolis (130). Since that time many modifications of this method has been used (i.e., Metropolis-Hastings (79) or Gibbs sampling (56)).

2.3.5 Diffusion approximation

Assume that we have a discrete Markovian process with the transition probability matrix $P = \{p_{ij}\}$ and transition probability densities given by $Q(t) = \{q_{ij}(t)\}$. Then, we can approximate this process by a continuous-time process using diffusion approximation. The prototype for diffusion processes is the Brownian motion (or the Wiener process) (147). It is a process with normally distributed increments. Einstein showed (39) that the densities of this process satisfy the following equation:

$$\frac{\partial q_{ij}(t)}{\partial t} = \frac{1}{2} b \frac{\partial^2 q_{ij}(t)}{\partial i^2}, \tag{2.8}$$

where bt is a variance of the process. We introduce the general discrete-jump process and postulate that the moments of the change δi, given the current value i at time t, satisfy the equations (44, 49):

$$\mathrm{E}(\delta i) = a(i)\delta t + o(\delta t) \tag{2.9}$$

$$\mathrm{Var}(\delta i) = b(i)\delta t + o(\delta t) \tag{2.10}$$

$$\mathrm{E}(|\delta i|^3) = o(\delta t) \tag{2.11}$$

,

where a and b are functions of i, often identified as infinitesimal velocity (a) or infinitesimal variance (b). It leads to the forward Kolmogorov (or Fokker-Planck) equation (44):

$$\frac{\partial q_{\cdot i}(t)}{\partial t} = -\frac{\partial}{\partial i}\left(a(i) q_{\cdot i}(t)\right) + \frac{1}{2}\frac{\partial^2}{\partial i^2}\left(b(i) q_{\cdot i}(t)\right) \tag{2.12}$$

If we denote the initial value of the diffusion variable by j, then we obtain the backward Kolmogorow equation (44):

$$\frac{\partial q_{ji}(t)}{\partial t} = a(j)\frac{\partial q_{ji}(t)}{\partial j} + \frac{1}{2} b(j) \frac{\partial^2 q_{ji}(t)}{\partial j^2} \tag{2.13}$$

One may usually obtain both functions a and b from the process and use either Formula (2.12) or Formula (2.13) to approximate the discrete process. Using continuous-time approach often decreases the complexity of the calculations, but one need to be aware that, in some cases (i.e., when the population size is small), the diffusion approximation may cause significant accuracy errors.

3

Problem statement

3.1 Aim of the dissertation

In the previous chapter of the dissertation we have shortly described the main approaches used in modeling of the evolution of genetic systems. Most of the recently developed theoretical stochastic models are either very simplified or too complex to obtain any exact results leading to the necessity of applying effective and sophisticated computer simulations or heuristic algorithms. A significant majority of these methods rely on the statistical analysis of real or artificially generated (sampled) data. The Monte Carlo based approaches are commonly used in order to obtain averaged results of simulations. The quality of the results obtained from mentioned methods strongly depends on the quantity and the quality of the data and is a compromise between the complexity of the model and the time spent on simulations. The non-simulation approaches have similar, but little different in details, limits. In both cases we cannot improve the data we are working on, but the time necessary to obtain the results that satisfy our quality requirements depends on different factors. The simulation-based methods allow to build much more complex models but require many single experiments to be carried out, often with additional heuristics involved.

Other important issue with the simulation-based methods is that they require much more careful verification. The simulations are usually used for the cases that are impossible to analyze by any analytical approaches. Therefore, the simplified non-simulation methods may deliver a test platform for these methods.

In this dissertation we focus on the non-simulation based genetic systems that model

non-trivial aspects of the evolution. These systems, in order to generate any useful and interesting results, require often an application of very complex calculations, impossible to solve by any analytical methods. **We want to show that these systems can be successfully used if they are refined by sophisticated, specially developed computer algorithms.** Moreover, the genetic systems built in such a way can, for specific problems, achieve better results (faster obtained and more accurate) than simulation-based approaches. We will present (in Chapters 4-6) three different models realizing such systems. In the following sections of this chapter we shortly discuss the reasoning behind each model. Although all of the presented models conduct the exact calculations, we will also use simulation methods, usually applied in order to compare both kinds of systems or to demonstrate sample results that can be obtained from our models. The complete list of algorithms is included at the end of the dissertation.

3.2 Moran model with recombination, mutation and drift

The studies of the interactions between mutation, drift and recombination has been dramatically widened in the recent years. Most of these studies explain the effect of recombination in the context of the backward-time coalescent theory (8, 38, 66, 90, 179). The backward-time approach, much faster than the forward-time solution, does not allow to model accurately more complex recombination aspects, especially if a system with multiple linked loci is considered. Thus, with the dramatic increase of the computer power, the forward-time approaches become currently feasible to model recombination and many forward-time (11, 71, 83, 87, 143, 152) or mixed (146) simulation packages have been developed in the last decade. However, there are some aspects of the population genetics under recombination that are still to be clarified. One of them is the question of the identifiability, i.e., if the population can reach a stage at which it is indistinguishable from the population evolving solely under drift and mutation. In this dissertation we prove that **it is possible, using a non-simulation approach applied to the mathematical Moran model, to answer the question of the recombination identifiability, at least in the means of the relationships limited to a set of distributions, which jointly characterize allelic states at a number of different loci.**

3.3 Genetic drift model in populations with time-varying size

The knowledge of the exact or the distribution of the time to the most recent common ancestor of a given population provides us with information about evolutionary history of this population. We can also use this knowledge to estimate other significant parameters of the population. As an example, the time to the MRCA is closely related to the relatedness of sampled individuals. Finding of the time to the MRCA of a subpopulation may also be useful in analyzing the whole population. Thus, determining the time to the MRCA has been under extensive mathematical analysis since a long time (65, 110, 120). Many aspects of the calculation of the time to the MRCA for simple models have been studied in details and are very well known (179). In the simple coalescent models, especially based on the Wright-Fisher model, the population size either is constant or it changes over time according to a specific growth rate (usually exponential). Recently, more complex models have been used in investigating the time to the MRCA. As an example, we can mention diffusion methods applied to the Wright-Fisher models (155, 167) or to the branching models (43). An interesting problem is the study of a population without assuming any specific population growth model. In this dissertation we provide a method which shows that **it is possible, using a recursive algorithm, to calculate the exact distribution of the time to the MRCA of a large sample from a population evolved under any growth scenario with the time efficiency of the method allowing for analysis of large human populations.** The change of the population size over time is the only information necessary to apply our model.

3.4 Demographic network

Traits of interactions between many sets of different species or different populations (including extinct ones) in the same species can be found in the genoms of their individuals. The reasons for studying of these interactions vary from the curiosity about common history of these species/populations to more important reasons providing insights into the genealogy of mutations that may be used in the development of gene mapping of rare (Mendelian) disease mutations methods (186). However, explaining all

of details of these interactions based on a given sample (usually containing not enough number of individuals) is not an easy task. Available approaches (72) try to estimate parameters that describe these interactions by simulating samples that fits the given data set. Demography is usually assumed by scientists and tested by the comparison of the obtained by simulations results to the data derived from real samples.

We claim that **it is possible to build a non-simulation model of demographic interactions between many populations or species that can, in some applications, replace the simulation-based approach.** In this dissertation we introduce a method to model such a complex demography network. Besides the demography events (such as a split of a single population into two populations, a merge of two populations or a migration event between populations), the basic version of our approach can describe the genetic drift inside the population, change of the population size over time and any Markovian-like mutation model. This model can also be refined by introducing other evolutionary forces. The result of our model is a joint distribution of pairs of individuals sampled from two populations (possibly the same). Based on the knowledge of the exact values of this distribution one can easily obtain other interesting parameters describing interactions between modeled populations (i.e., ascertainment bias (177) or pairwise difference (24)). The main advantages of our approach are: (i) capability to model a very complex demography in efficient way (ii) obtaining the exact values of the distribution without necessity of simulations and (iii) capability to model a great variety of population models.

4

Moran model with drift, mutation and recombination

4.1 Mathematical preliminaries

4.1.1 Markov semigroup theory

Distribution

We assume that $L^1(\mathbb{X}, \sum, f)$ stands for the space of absolutely summable sequences $(\xi_i)_{i\in\mathbb{A}}$, with the norm $\|(\xi_i)_{i\in\mathbb{A}}\| = \sum_{i\in\mathbb{A}} |\xi_i|$. The elements $e_j = (\delta_{ij})_{i\in\mathbb{A}}$, $j \in \mathbb{A}$, where δ_{ij} is the Kronecker delta, form the basis of L^1. Any $(\xi_i)_{i\in\mathbb{A}} \in L^1$ may be represented as $(\xi_i)_{i\in\mathbb{A}} = \sum_{i\in\mathbb{A}} \xi_i e_i$. We define the vector $(\xi_i)_{i\in\mathbb{A}} \in L^1$ as a distribution if and only if $\xi_i \geq 0, i \in \mathbb{A}$, and $\sum_{i\in\mathbb{A}} \xi_i = 1$.

$\mathcal{M}_n$, where n is an integer, is the space of absolutely summable n-dimensional matrices $\mathrm{m} = (a_{i_1,\ldots,i_n})_{i_1,\ldots,i_n\in\mathbb{A}}$ with the norm $\sum_{i_1,\ldots,i_n\in\mathbb{A}} |a_{i_1,\ldots,i_n}|$. We term the matrix $\mathrm{m} \in \mathcal{M}_n$ as a distribution if and only if its entries are non-negative and add up to 1. The distributions in $\mathcal{M}_n$ are distributions of n-tuples of $\mathbb{A}$-valued random variables. Thus, $\mathcal{M}_n$ is a tensor product of n copies of $L^1 : \mathcal{M}_n = (L^1)^{n\otimes}$.

Strongly continuous semigroup and its generator

A family of operators $\{S(t), t \geq 0\}$ in a Banach space $\mathbb{X}$ is called a strongly continuous semigroup (42) if and only if satisfies the following conditions for all $x \in \mathbb{X}$ (163):

- $S(0)$ is the identity operator

- $S(t)S(s) = S(t+s)$
- $\lim_{t\to 0} S(t)x = x$

The infinitesimal generator G of $S(t)$ is defined by:

$$Gx = \lim_{t\to 0} \frac{1}{t}(S(t) - I)x \tag{4.1}$$

whenever the limit exists. The domain of G is the set of $x \in X$ for which this limit exist. The domain is a linear subspace and G is linear on this domain (149). The generator G is the expected infinitesimal change of x applied onto process described by semigroup. Thus, the infinitesimal generator is the measure of behavior of the process.

Markov semigroup

A strongly continuous semigroup in L^1 is termed a Markov semigroup if and only if all $S(t)$ operators are Markov operators. A Markov operator is an operator that maps distribution into distribution. Thus, $\|S(t)\| = 1$. Markov operators are usually given by a transition probability function (82).

If $\{S_i(t), t \geq 0\}$, $i = 1, \ldots, n$ are strongly continuous semigroups of Markov operators in L^1, then $\{S_1(t) \otimes \cdots \otimes S_n(t), t \geq 0\}$ is a strongly continuous Markov semigroup in $\mathcal{M}_n$ and it is the tensor product of $\{S_i(t), t \geq 0\}$, $i = 1, \ldots, n$.

The Cartesian product $\mathcal{M}_n^m$ of m copies of $\mathcal{M}_n$ provides a way of gathering information on distributions of m n-tuples of $\mathbb{A}$-valued variables. We may see this space as a direct sum of m copies of $\mathcal{M}_n$. We say that $x \in \mathcal{M}_n^m$ is a distribution if and only if it is a convex combination of m distributions in $\mathcal{M}_n$. Any operator in $\mathcal{M}_n^m$ is a Markov operator mapping distributions into distributions. If $\{T_i(t), t \geq 0\}$, $i = 1, \ldots, m$ are Markov semigroups in $\mathcal{M}_n$, then $\{\bigoplus_{i=1}^m T_i(t), t \geq 0\}$, defined as $\bigoplus_{i=1}^m T_i(t)(\sum_{i=1}^m m_i) = \sum_{i=1}^m T_i(t)m_i$ is a Markov semigroup in $\mathcal{M}_n^m$. The domain of the infinitesimal generator $\mathcal{G}$ of this semigroup is the Cartesian product of domains of generators G_i of $\{T_i(t), t \geq 0\}$ and we have $\mathcal{G}(m_i)_{i=1,\ldots,m} = (G_i m_i)_{i=1,\ldots,m}$ for $(m_i)_{i=1,\ldots,m}$ in this domain.

4.1.2 Markov chains

Ergodicity

Irreducible Markov chain is a Markov chain in which for each pair of states (i, j) the following condition is satisfied:

$$\exists_{k\geq 1} P^k(i,j) > 0 \tag{4.2}$$

where $P(i, j)$ is the entry of the transition probability matrix of the given Markov chain describing the probability of moving from state i to state j.

Irreducible Markov chain is either periodic or ergodic. If the chain is periodic, then for each pair of states i, j exists period $k \geq 2$ that $P^k(i, j) = P(i, j)$. The values of the period for each pair of states in periodic Markov chain are the same. If all entries $P(i, j)$ of an irreducible Markov chain are positive, then this Markov chain is ergodic.

If an irreducible Markov chain with transition probability matrix P and k states is aperiodic (ergodic), then there exists a vector $\pi : \{\pi_i\}$, $i = 1, \ldots, k$, called as a stationary or an equilibrium distribution, that satisfies the following conditions:

- $\pi_i > 0$, where $1 \leq i \leq k$
- $\sum_{i=1,\ldots,k} \pi_i = 1$
- $\forall_{1\leq i,j\leq k} \lim_{x\to\infty} P^x(i, j) = \pi_j$

Dobrušin's coefficient

For a Markov chain with the transition probability matrix $P = \{P(i, j)\}$, where $1 \leq i, j \leq k$, we define the Dobrušin's coefficient of ergodicity β (37) given by the following formula:

$$\beta = \beta(P) = \min_{1\leq i,j\leq k} \sum_{m=1}^{k} \min(P(i,m), P(j,m)) \tag{4.3}$$

If the value of the Dobrušin's coefficient $\beta(P) > 0$, then the given Markov chain is ergodic and the value of the coefficient describes the speed of convergence to the stationary distribution (the greater the value, the faster it converges):

$$\|P^x - \Pi\| \leq (1 - \beta(P))^x \tag{4.4}$$

where $\|\cdot\|$ denotes the maximum of all absolute values of entries of the matrix.

Spectral theory

By definition (165), the spectral gap of the transition probability matrix P of a Markov chain is equal to the smallest nonzero eigenvalue of the matrix $Q = I - \frac{1}{2}(P + P^*)$, where I is an identity matrix and P^* is the transition matrix of the time-reversed process. Each entry of the matrix $P^*(i,j)$ is defined as: $P^*(i,j) = P(j,i)\frac{\pi_j}{\pi_i}$, where π is the stationary distribution of the matrix P.

Analysis of the values of the spectral gap allow us to obtain information about behavior of the process described by the given Markov chain. The higher value of the spectral gap is indicative of faster convergence to the equilibrium (134).

4.2 Generalization of 2 loci Kimmel-Polańska-Bobrowski model

The model of our interest was firstly introduced by Kimmel and Polańska (108) to model pairs of repeat-DNA sequences (microsatellites). The model included recombination between two loci along with genetic drift and mutations. More aspects of this simple version of the model, regarding asymptotic behavior under different demographic scenarios, were discussed in (18). However, all these studies were conducted only for the most basic model of recombination and cannot be generalized for a model with greater number of loci.

4.2.1 Model description

Mutation model

We assume that modeled population is composed of $2N$ individuals. Each individual consists s loci which leads to $s-1$ possible recombination sites. Individuals are represented as s-tuples of $\mathbb{A}$-values random variables $(X_{a,b})_{1\leq a\leq 2N, 1\leq b\leq s}$, where a is the individual number and b is the index of locus on a chromosome. We assume that these tuples are exchangeable and that each of them evolves in time as non-explosive Markov chains, independent of the other ones, but with the same transition probabilities. This models mutation at all loci of a chromosome in each individual. The process of mutation at the locus b in each individual is described by means of a strongly continuous semigroup $\{S_{X_b}(t), t \geq 0\}$ of Markov operators in L^1. This means that if $x \in L^1$ is

the distribution of allele types at time 0 then $S_{X_b}(t)x$ is the distribution of allele types at time t. The tensor product semigroup $\{S(t), t \geq 0\}$, $S(t) = S_{X_1}(t) \otimes \cdots \otimes S_{X_s}(t)$, describes evolution of distributions at s loci, provided mutations at these loci occur independently.

Mating scheme

We incorporate recombination and genetic drift in the model by assuming that each individual's life-length is an exponential random variable with a parameter $\frac{2}{\lambda}$ and that at the moment of individual's death, the s-tuple by which it is represented is replaced by another s-tuple in the following manner. Three numbers j, k, m are randomly chosen with replacement from $1, 2, \ldots, 2N$. The j value indicates the deceased individual being the s-tuple $(X_{j1}, X_{j2}, \ldots, X_{js})$, $j = 1, \ldots, 2N$. With probability $1 - r$, where $r = \sum_{i=1}^{s-1} r_i$ with $r_i \in [0, 1]$ such that $r \in [0, 1]$, being the given parameters, one of the s-tuples $(X_{k1}, X_{k2}, \ldots, X_{ks})$, $k = 1, \ldots, 2N$ replaces the deceased one. With probability r_i, $1 \leq i \leq s - 1$, the recombination event occurs after the ith locus. In this case the i-tuple $(X_{k1}, \ldots, X_{ki})$, $k = 1, \ldots, 2N$ is drawn at random first and the $(s - i)$-tuple $(X_{m(i+1)}, \ldots, X_{ms})$, $m = 1, \ldots, 2N$ is drawn next independently from the first draw. As a consequence, a new s-tuple becomes one of the already existing s-tuples (including the one just deceased) $(X_{j1}, \ldots, X_{js})$, each of them with probability $(1 - r)\frac{1}{2N} + r\frac{1}{(2N)^{s-1}}$, or one of the $s - 1$ types of "mixed ones": either $(X_{j1}, \ldots, X_{ji}, X_{k(i+1)}, \ldots, X_{ks})$, $j \neq k, 1 < i \leq s$ each with probability $r_i \frac{1}{(2N)^{s-1}}$.

If the s-tuples $(X_{i1}, \ldots, X_{is})$ are exchangeable, then, because of the sampling scheme, it is obvious that so are the newly formed s-tuples immediately after individual's death. This fact follows from Lemma 1 in (18) if we note that for the recombination after the jth locus $(1 \leq j < s)$ both: the j-tuple $(X_{i1}, \ldots, X_{ij})$ and the $(s - i)$-tuple $(X_{i(j+1)}, \ldots, X_{is})$ can be treated as a single compound locus. Therefore, exchangeability is preserved in the model.

Distribution's relations on the individual's death

Let $(X_{a1}, \ldots, X_{as})$ and $(\tilde{X}_{a1}, \ldots, \tilde{X}_{as}), a = 1, \ldots, 2N$ be the s-tuples representing individuals in the population immediately before and immediately after individual's death. The total number of tuples is equal to $(2N)^s$ which is large enough to make any analysis of the population impossible. Fortunately, based on exchangeability, we can reduce the

number of states in our model. By exchangeability of $(X_{a1}, \ldots, X_{as})$, $a \in 1, \ldots, 2N$, the distribution of $(X_{a_1 1}, \ldots, X_{a_s s})$ where $1 \leq a_i \leq 2N$ does not depend on the choice of a_i but only on the mutual relations between a_i values (their equalities). For example the distribution of (X_{51}, X_{32}, X_{53}) (where first and third loci descended from the same individual) is exactly the same as (X_{21}, X_{12}, X_{23}) or any (X_{a1}, X_{b2}, X_{a3}) where $1 \leq a, b \leq 2N, a \neq b$. We will denote these new distributions by $D_{a_1 \ldots a_s}$. The corresponding Ds with tilde denote distributions in the population immediately after individual's death. Equality of $a_i = a_j$ in the index of D means that the ith and the jth loci descended from the same individual from the first generation.

To order all $D_{a_1 \ldots a_s}$ distributions involved in the model we introduce multi-indexes $(a_1 \ldots a_s)$ that satisfies the following properties:

1. a_1 is 1,
2. $a_\alpha \leq \max(a_1, \ldots, a_{\alpha-1}) + 1$, $\alpha \geq 2$;

We call such multi-indexes regular. The first distribution of the regular multi-index is $D_{11\ldots1}$ and the last one is $D_{123\ldots s}$. There are ϖ_s regular multi-indexes, where ϖ_s is the Bell number, the number of ways to partition a set of s elements into subsets (63). For, with every partition we have a natural order of its elements (subsets) where the first subset is the one containing element 1 and the kth is the one containing the smallest number not included in the previous $k-1$ subsets (provided such number exists). To such naturally ordered partition we assign a regular multi-index by labeling elements of the kth subset with label k, and this map is injective. On the other hand, given a regular multi-index, we obtain a partition by collecting all numbers with the same index into one subset. Such an assignment of partition is injective, since the multi-index agrees with the labeling obtained from the natural order.

Finally, we arrange all distributions $D_{a_1 \ldots a_s}$ in lexical order, thus forming vector D. Similarly, we form the vector $\tilde{D}$ of distributions $\tilde{D}_{a_1, \ldots, a_s}$, and consider the way coalescence/recombination event influences it. Suppose the recombination occurred after the ith locus, we are interested in $\tilde{D}_{a_1, \ldots, a_s}$ and we know that the jth individual died to be replaced partly by the kth and partly by the mth. Then, $\tilde{D}_{a_1 \ldots a_s}$ equals $D_{k_1 \ldots k_s}$ where the multi-index $(k_1, \ldots, k_s)$ is formed as follows. First, all occurrences of j at up to and including ith place in $(a_1, \ldots, a_s)$ are replaced by k, and all the remaining

occurrences are replaced by m. Then, the newly formed multi-index is transformed into a regular multi-index as follows. First, we change all occurrences of a_1 to 1, if the first condition of regularity is not yet met. Next, we look for the first place, say a_α, where the second requirement is not met. If there is no such place, we are done. Otherwise, we replace a_α and all its occurrences by the smallest integer larger than all a_β preceding a_α, and we continue this procedure until the multi-index is regular.

As a result, $\tilde{D}_{a_1\ldots a_s}$ is a convex combination of all possible $D_{k_1\ldots k_n}$'s. Each choice of j, k and m adds the term $\frac{1}{(2N)^3} D_{k_1\ldots k_s}$ to this combination (all choices of j, k and m are equally likely). All coefficients of this combination are themselves linear combinations of $1, b, b^2$ and b^3 where $b = (2N)^{-1}$.

Hence, there exists a $\varpi_n \times \varpi_n$ transition matrix Θ of a Markov chain such that

$$\tilde{D} = \Theta D. \tag{4.5}$$

The Θ matrix is a convex combination of s transition matrices, corresponding to the cases of no recombination (Θ_0) and of recombination events after the ith locus ($\Theta_i, 1 \leq i < s$):

$$\Theta = (1-r)\Theta_0 + \sum_{i=1}^{s-1} r_i \Theta_i \tag{4.6}$$

4.2.2 Simple cases

In this section we present an exact value of the Θ matrix for the simplest cases ($s \leq 3$), along with the explanation of how these values were obtained. Since ϖ_s is a fast growing sequence ((63) p.693, e.g. $\varpi_4 = 15, \varpi_5 = 52$, and $\varpi_9 = 21147$), finding an explicit form of Θ for $s \geq 4$ can be done only by using of the computer program.

We assume that the death/birth process is described by three numbers $1 \leq j, k, m \leq 2N$ (individual j dies and is replaced by individual k recombining, with probability r, with individual m). To obtain the distribution $D_{b_1\ldots b_s}$ from which the distribution $\tilde{D}_{a_1\ldots a_s}$ was obtained under the given (j, k, m) triple, firstly we need to change the values of all $a_i = j, 1 \leq i \leq s$ into: (i) k if there is no recombination or ith locus is located before recombination site or (ii) m otherwise. After these changes, the newly formed multi-index may not be regular and thus, we need to restore his regularity obtaining the final $(b_1, \ldots, b_s)$ index.

Two loci model

In this case $\Theta = (1-r)\Theta_0 + r\Theta_1$. Since $\varpi_2 = 2$, we have only two possible distributions D_{11} and D_{12}. The first distribution describes individuals with both loci descended from the same individual and the second distribution describes all other cases. If there is no recombination, the only possibility to change the distribution in the death/birth process occurs when either $j = 1, k = 2$ or $j = 2, k = 1$. In this case $\tilde{D}_{12} = D_{11}$ leading to:

$$\Theta_0 = \begin{pmatrix} 1 & 0 \\ \frac{2}{(2N)^2} & 1 - \frac{2}{(2N)^2} \end{pmatrix} \tag{4.7}$$

When the recombination occurs, it is possible that $\tilde{D}_{11} = D_{12}$ (if $j = 1$ and $k \neq m$) or $\tilde{D}_{12} = D_{11}$ (if $j = 1$ and $k = 2$ or $j = 2$ and $m = 1$). This leads to:

$$\Theta_1 = \begin{pmatrix} 1 - \frac{2N-1}{(2N)^2} & \frac{2N-1}{(2N)^2} \\ \frac{2}{(2N)^2} & 1 - \frac{2}{(2N)^2} \end{pmatrix} \tag{4.8}$$

Finally, we get:

$$\Theta = \begin{pmatrix} 1 - r\frac{2N-1}{(2N)^2} & r\frac{2N-1}{(2N)^2} \\ \frac{2}{(2N)^2} & 1 - \frac{2}{(2N)^2} \end{pmatrix} \tag{4.9}$$

Three loci model

In this case $\Theta = (1-r)\Theta_0 + r_1\Theta_1 + r_2\Theta_2$. Vector D contains $\varpi_3 = 5$ distributions listed in lexical order: $D = [D_{111}, D_{112}, D_{121}, D_{122}, D_{123}]^T$.

If there is no recombination, none of D_{112}, D_{121} and D_{122} has changed unless $j = 1, k = 2$ or $j = 2, k = 1$. In this last case $\tilde{D}_{112} = \tilde{D}_{121} = \tilde{D}_{122} = D_{111}$. Hence,

$$\Theta_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ \frac{2}{(2N)^2} & 1 - \frac{2}{(2N)^2} & 0 & 0 & 0 \\ \frac{2}{(2N)^2} & 0 & 1 - \frac{2}{(2N)^2} & 0 & 0 \\ \frac{2}{(2N)^2} & 0 & 0 & 1 - \frac{2}{(2N)^2} & 0 \\ 0 & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & 1 - \frac{6}{(2N)^2} \end{pmatrix} \tag{4.10}$$

where the form of the last row is justified as follows. If $j = 1, k = 2$, then $(\tilde{X}_{11}, \tilde{X}_{22}, \tilde{X}_{33}) = (X_{21}, X_{22}, X_{33})$ and so $\tilde{D}_{123} = D_{112}$. Similarly we show that this equality is true when $j = 2$ and $k = 1$. Analogously, $\tilde{D}_{123} = D_{122}$ if either $j = 2, k = 3$ or $j = 3, k = 2$, and $\tilde{D}_{123} = D_{121}$ if either $j = 3, k = 1$ or $j = 1, k = 3$. In the remaining cases $\tilde{D}_{123} = D_{123}$.

Table 4.1: Calculation of Θ_1. If $j = 1$ and recombination took place after the first locus, $(\tilde{X}_{11}, \tilde{X}_{12}, \tilde{X}_{13}) = (X_{k1}, X_{m2}, X_{m3})$ and so $(\tilde{X}_{11}, \tilde{X}_{22}, \tilde{X}_{13}) = (X_{k1}, X_{22}, X_{m3})$. Considering all possible cases for k and m we obtain four entries in the middle row of the table. The remaining entries in the table are obtained similarly.

	$j \neq 1, 2$	$j = 1,$ $k \neq 2,$ $m \neq 2$	$j = 1,$ $k = 2,$ $m \neq 2$	$j = 1,$ $k = 2,$ $m = 2$	$j = 1,$ $k \neq 2,$ $m = 2$	$j = 2,$ $m \neq 1$	$j = 2,$ $m = 1$
$\tilde{D}_{112}$	D_{112}	$D_{112}, k = m$ $D_{123}, k \neq m$	D_{121}	D_{111}	D_{122}	D_{112}	D_{111}
$\tilde{D}_{121}$	D_{121}	$D_{121}, k = m$ $D_{123}, k \neq m$	D_{112}	D_{111}	D_{122}	D_{121}	D_{111}
$\tilde{D}_{122}$	D_{122}	D_{122}	D_{111}	D_{111}	D_{122}	D_{122}	D_{111}

To find the three rows in the middle of Θ_1 we consider recombination between the first two loci by listing the possible cases in Table 4.1. This gives Θ_1 in the form:

$$\begin{pmatrix} 1-\frac{2N-1}{(2N)^2} & 0 & 0 & \frac{2N-1}{(2N)^2} & 0 \\ \frac{2N+1}{(2N)^3} & \frac{2N-2}{2N}+\frac{2N-1}{(2N)^3}+\frac{2N-1}{(2N)^2} & \frac{2N-1}{(2N)^3} & \frac{2N-1}{(2N)^3} & \frac{(2N-1)(2N-2)}{(2N)^3} \\ \frac{2N+1}{(2N)^3} & \frac{2N-1}{(2N)^3} & \frac{2N-2}{2N}+\frac{2N-1}{(2N)^3}+\frac{2N-1}{(2N)^2} & \frac{2N-1}{(2N)^3} & \frac{(2N-1)(2N-2)}{(2N)^3} \\ \frac{2}{(2N)^2} & 0 & 0 & 1-\frac{2}{(2N)^2} & 0 \\ 0 & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & \frac{2}{(2N)^2} & 1-\frac{6}{(2N)^2} \end{pmatrix}. \tag{4.11}$$

Obtaining the first row here is straightforward (similarly to two loci case) , and the last row is obtained by noting that:

- for $j = 1$, $\tilde{D}_{123} = D_{112}$ provided that $k = 2$, $\tilde{D}_{123} = D_{121}$ provided that $k = 3$, and $\tilde{D}_{123} = D_{123}$ in the remaining cases
- for $j = 2$, $\tilde{D}_{123} = D_{112}$ provided that $m = 1$, $\tilde{D}_{123} = D_{122}$ provided that $m = 3$, and $\tilde{D}_{123} = D_{123}$ in the remaining cases
- for $j = 3$, $\tilde{D}_{123} = D_{121}$ provided that $m = 1$, $\tilde{D}_{123} = D_{122}$ provided that $m = 2$, and $\tilde{D}_{123} = D_{123}$ in the remaining cases
- for $j \geq 4$, $\tilde{D}_{123} = D_{123}$

To cover the case of recombination after the second locus we note that our model is symmetric with respect to numbering loci. More specifically, if the loci were numbered

from the last one to the first, the distributions D_{111}, D_{112}, $D_{121}, D_{122}, D_{123}$ would have become D_{111}, D_{122}, $D_{121}, D_{112}, D_{123}$, which amounts to transposition of D_{112} and D_{122}. Since such a numbering transposes recombination loci, the following lemma is correct:

Lemma 4.1. *For $s = 3$ the matrix Θ_2 may be obtained directly from Θ_1 by interchanging columns 2 and 4 and, next, interchanging rows 2 and 4.*

4.3 Asymptotic behavior

4.3.1 The general case

All distributions D form a complete system in that their evolution in time depends merely on their initial conditions, matrix Θ and the semigroup $\{S(t), t \geq 0\}$ in $\mathcal{M}_s$. For, if we let G be the generator of $\{S(t), t \geq 0\}$ and $\mathcal{G}$ be the generator of the Cartesian product $\{\mathcal{S}(t), t \geq 0\}$ of ϖ_s copies of $\{S(t), t \geq 0\}$ in $\mathcal{M}_s^{\varpi_s}$, then writing $D(t)$ for the (column-) vector of D we have

$$\frac{\mathrm{d}D(t)}{\mathrm{d}t} = \mathcal{G}D(t) + \lambda N\Theta D(t) - \lambda N D(t), t \geq 0 \tag{4.12}$$

provided that $D(0)$, the initial state of the distributions, belongs to $\mathcal{D}(\mathcal{G})$, for example if all of its coordinates belong to $\mathcal{D}(G)$. In other words, $D(t) = \mathcal{T}(t)D(0)$ where the semigroup $\{\mathcal{T}(t), t \geq 0\}$ is generated by $\mathcal{G} + \lambda N\Theta - \lambda N$. The proof of these facts is analogous to that given in (18) where the case of two loci is treated. Moreover, Θ commuting with $\mathcal{S}(t)$ gives us following formula (64):

$$\mathcal{T}(t) = \mathcal{S}(t)\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} \tag{4.13}$$

The result is intuitively clear: in the absence of genetic drift, where the members of the population evolve without influencing one another, the behavior of D is governed by (4.12) with $\lambda = 0$. In this case (4.12) is an uncoupled system of ϖ_s independent equations ($\lambda = 0$ gives infinite life-time of an individual). The process of birth-death events is then treated as a perturbation of the uncoupled system and these events occur at an exponential rate λN (since there are $2N$ individuals, each of them having independent exponentially distributed life-lengths with parameter $\frac{2}{\lambda}$).

Following lemma determines the ergodicity of the matrix Θ.

Lemma 4.2. *For any number of loci the transition matrix $\Theta = (\theta_{ij})$ where i, j are in $\{1, \ldots, \varpi_s\}$ is ergodic.*

Proof. The Dobrušin's coefficient value β of the matrix Θ is equal to 0 if $s \geq 4$ and thus it cannot be used directly to prove the ergodicity of the matrix. If we treat $2s-2$ steps of the Markov chain described by Θ as a single step, we get a Markov chain with transition probability matrix $\Theta^{(2s-2)}$. The egrodicity of this new Markov chain follows with the ergodicity of the basic Markov chain. Two facts are true:

- It is possible to move from the $1,\ldots,1$ state to any other state in at most $s-1$ steps. To prove this, let $a_1,\ldots,a_s$ be an arbitrary regular multi-index. Consider a recombination event: let i be the recombination site number, j be the number of an individual to be replaced, k be the number of an individual supplying the loci with numbers 1 through i, and m be the number of an individual supplying the loci with numbers $i+1$ through s. Taking $i=1, j=k=1$ and $m=a_2$ we jump from $(1,\ldots,1)$ to $(a_1,a_2,\ldots,a_2)$, a_1 equaling 1 by assumption. After arriving at $(a_1,a_2,\ldots,a_b,\ldots,a_b), b \geq 2$ we choose $i=l, j=k=a_b$, $m=a_{b+1}$ to jump to $(a_1,a_2,\ldots,a_b,a_{b+1},\ldots,a_{b+1})$. Hence, after $s-1$ jumps, we arrive at $a_1,\ldots,a_s$.

- It is possible to move from any $a_1,\ldots,a_s$ state to the $1,\ldots,1$ state in at most $s-1$ steps. Starting from $a_1,\ldots,a_s$, we choose $i=s-1, j=k=a_s, m=a_{s-1}$ to jump to $(a_1,\ldots,a_{s-2},a_{s-1},a_{s-1})$. After arriving at $(a_1,\ldots,a_{s-b-1},a_{s-b},\ldots,a_{s-b})$, $1 \leq b \leq s-2$ we choose $i=s-b-1, j=k=a_{s-b}$ and $m=a_{s-b-1}$, to jump to $(a_1,\ldots,a_{s-b-2},a_{s-b-1},\ldots,a_{s-b-1})$. Hence, after $s-1$ jumps we arrive at $(a_1,\ldots,a_1)=(1,\ldots,1)$,.

These facts show that any two states $1 \leq i,j \leq \varpi_s$ communicate with each other in at most $2s-2$ steps (by moving from ith state to $1,\ldots,1$ firstly and then to jth state). Thus $\theta_{ij}^{(2s-2)} > 0$ for each $1 \leq i,j \leq \varpi_s$ leading to $\beta(\Theta^{(2s-2)}) > 0$ which proves our claim. □

Based on the ergodicity of the transition probability matrix Θ, we define Π as a $\varpi_s \times \varpi_s$ matrix with all rows equal to a stationary distribution of the matrix Θ such that $\lim_{t\to\infty} \|\mathrm{e}^{-\lambda N t}\mathrm{e}^{Nt\Theta} - \Pi\| = 0$.

As a result we obtain Theorem (4.1)

Theorem 4.1. *For any number of loci* $\lim_{t\to\infty} \|\mathfrak{T}(t)D(0) - \mathfrak{S}(t)\Pi D(0)\| = 0$.

The most interesting practical consequence of Theorem 4.1 is that for large t, the distribution $D_{1\ldots1}(t)$ in the model with drift and recombination is asymptotically the same as that in the model without drift and recombination provided that initial condition

in the latter is the appropriate convex combination involving stationary distribution of the matrix Θ:

$$D_1(t) \sim S(t) \sum_{\iota=1}^{\varpi_s} \pi_\iota D_\iota(0),$$

where instead of $D_{a_1 \ldots a_s}$ we write D_ι where $\iota = \iota(a_1, \ldots, a_s)$ denotes the position of $D_{a_1 \ldots a_s}$ in D (for example $\iota(1, \ldots, 1) = 1$ and $\iota(1, 2, \ldots, s) = \varpi_s$). In other words, recombination influences the model merely through this stationary distribution and this is regardless of the way mutations are modeled.

4.3.2 Simple case – three loci model

For the model with three loci both, Lemma 4.2 and Theorem 4.1 apply but the knowledge of the exact forms of matrices $\Theta_i, 0 \leq i \leq 2$ allows us to formulate more accurate Theorem 4.2.

Theorem 4.2. *For the model with three loci* $\lim_{t\to\infty} \|\mathcal{T}(t)D(0) - \mathcal{S}(t)\Pi D(0)\| = 0$ *and the speed of convergence is exponential.*

Proof. To estimate the Dobruš in's coefficient of Θ we note that $\alpha = 2-2\beta$ (which is the maximum appearing in (4.3)) is a convex function of Θ, and so β is concave. Therefore, $\beta(\Theta) \geq (1-r)\beta(\Theta_0) + r_1\beta(\Theta_1) + r_2\beta(\Theta_2)$. Since, for Θ_0 this maximum is attained for rows $i = 1$ and $j = 5$, and equals 2, $\beta(\Theta_0) = 0$. Similarly, the maximum for Θ_1 is attained simultaneously for $(i, j) = (1, 2), (1, 3), (1, 5), (2, 4), (3, 4)$ and $(4, 5)$ (provided $2N \geq 3$), and equals $2 - \frac{1}{N^2}$. Hence, $\beta(\Theta_1) = \frac{2}{(2N)^2}$. Finally, since interchanging rows and columns does not influence the value of β and based on Lemma 4.1, $\beta(\Theta_2) = \beta(\Theta_1)$. Hence, $\beta(\Theta) \geq r\beta(\Theta_1) \geq \frac{2r}{(2N)^2} > 0$.

By (4.4) and (4.13) and $\Pi^x = \Pi$, $x \geq 2$, we have $\|\mathrm{e}^{\lambda Nt\Theta} - \mathrm{e}^{\lambda Nt\Pi}\| \leq \sum_{x=1}^{\infty} \frac{(\lambda Nt)^x \|\Theta^x - \Pi\|}{x!} \leq \sum_{x=1}^{\infty} \frac{(\lambda Nt)^x (1-\beta(\Theta))^x}{x!} \leq \mathrm{e}^{\lambda Nt(1-\beta)}$. Thus, $\|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} - \mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Pi}\| \leq \mathrm{e}^{-\lambda Nt\beta(\Theta)}$. Using $\mathrm{e}^{\lambda Nt\Pi} = I + \Pi(\mathrm{e}^{\lambda Nt} - 1)$, we get

$$\begin{aligned} \|\mathcal{T}(t) - \mathcal{S}(t)\Pi\| &\leq \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} - \Pi\| \\ &\leq \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Theta} - \mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Pi}\| + \|\mathrm{e}^{-\lambda Nt}\mathrm{e}^{\lambda Nt\Pi} - \Pi\| \\ &\leq \mathrm{e}^{-\lambda Nt\beta(\Theta)} + \mathrm{e}^{-\lambda Nt} \leq \mathrm{e}^{-\frac{\lambda r}{2N}t} + \mathrm{e}^{-\lambda Nt}, \end{aligned} \tag{4.14}$$

proving our claim. □

The knowledge of the exact forms of Θ_i, $0 \leq i \leq 2$ allows us also to find an explicit form of the stationary distribution π using *Mathematica* (4). Unfortunately,

the formula is long and non-informative. However, if we disregard all the terms of $\left(N^{-2}\right)$ order from Θ_1 and Θ_2, which for large populations are insignificant, and assume $r_1 = r_2$ the formula simplifies to give:

$$\pi_1 = a^2 \frac{b(3b-5) - 3a(2-3b+b^2)}{(ab-a-b)(6a^2+5ab-9a^2b+b^2-4ab^2+3a^2b^2)},$$
$$\pi_2 = \frac{ab(a-1)[3a(b-1)-2b]}{(ab-b-2a)(ab-a-b)(3ab-3a-b)},$$
$$\pi_3 = \frac{ab^2(a-1)}{3a^3(b-2)(b-1)^2 - b^3 + ab^2(5b-6) + a^2b(18b-7b^2-11)},$$
$$\pi_4 = \frac{ab(a-1)[3a(b-1)-2b]}{(ab-a-b)(6a^2+5ab-9a^2b+b^2-4ab^2+3a^2b^2)},$$
$$\pi_5 = \frac{(a-1)b^2(3a-1)}{ab(5-4b) + b^2 + 3a^2(2-3b+b^2)},$$

where $b = \frac{r}{2}$, and $a = \frac{1}{2N}$. In particular, if $\frac{1}{2N} \ll b$, $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) \approx (0, 0, 0, 0, 1)$ while if $2Nb \to c$, then $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5)$ is approximately equal to
$\frac{1}{(c+1)(c+2)(c+3)}(5c+6, c(3+2c), c^2, c(3+2c), c(c+1))$.

4.4 Model implementation

4.4.1 Computer programs

It was necessary to develop a computer program calculating Θ matrix in order to obtain any numerical results for $s \geq 4$. The program (named *theta*) is available on the disc attached to the paper. The program realizes following functionality:

- calculates the explicit form of the Θ_i, $1 \leq i \leq s$ matrices; these matrices are represented in symbolic way, each entry of the symbolic matrices is a triple of linear combination coefficients of a, a^2 and a^3, where $a = (2N)^{-1}$; these entries either correspond exactly to the values of real Θ_i probabilities (if they are not on the main diagonal) or require 1 to be added (otherwise – to the entries on the main diagonal)

- calculates the numerical values of the Θ_i, $1 \leq i \leq s$ matrices; to achieve this, the symbolic representation of the matrix along with the value of population size $2N$ is required

- calculates the stationary distribution of the Θ (with a given precision)

- calculates the Dobrušin coefficient, either the exact value (for the numerical matrix) or the sign (positive or equal to 0 – for the symbolic matrix)

Mentioned functionality is available through Recombination class (see attached resources for more details). Due to very large sizes of the Θ matrices even for small s values, the program handles only one symbolic and one numerical matrix at once (other matrices that are being used are kept as text files). The names of the result files containing Θ matrices starts with the number (being the number of loci s) and follows with a symbol of the matrix type ('S' – symbolic, 'V' – numerical), description of the recombination site (string 'main' – the main Θ matrix or string 'j..jk..k' of the s length where the j characters stands for the loci located on the left side of the recombination site and k characters otherwise) and optional number x used only for the numerical main matrix (it means that the presented matrix is equal to Θ^{2^x}). The program works well for $s \leq 9$ (for $s \geq 10$ the size of Θ makes the task unmanageable even for a computer).

We also attach two other computer programs. The first one, *spectralGap*, calculates the spectral gap of the given transition probability matrix according to the formulas presented in Section 4.1.2. To calculate the stationary distribution π, we apply the main program described above. Eigenvalues of the matrix Q are obtained by the computer program based on the QR algorithm (60).

The second program, *coalcov*, applies Monte-Carlo simulations to estimate the correlation of the time to the MRCA at two loci in a sample of two-loci individuals under a Moran model. The idea of the simulations is based on the Hudson's simulations (90).

4.4.2 Algorithms

Distribution's managing

We represent each of the possible ϖ_s distributions $D_{a_1 \ldots a_s}$ (with $a_1 \ldots a_s$ being the regular multi-index) as a string of length s with each character equal to $a_i + 48$, where 48 is an ASCII code of 0. Thus, for manageable cases $s \leq 9$, each a_i, $1 \leq i \leq s$ is a digit. We store all distributions in an array in the lexical order, starting from "$11 \ldots 1$" and ending with "$12 \ldots s$". To obtain such a set of distributions A we use Algorithm 4.1.

Algorithm 4.1 Representation of the distributions in the lexical order.

```
A ← 1 (the only distribution for s = 1)
for i = 2 to s
    B ← ∅
    foreach distribution D_{a1...a_{i-1}} ∈ A
        σ_i = max(a_1, ..., a_{i-1})
        for j = 1 to σ_i + 1
            B ← D_{a1...a_{i-1} j}
        endfor
    endfor
    A = B
endfor
```

We distinguish two basic operations on the distributions. The first one, termed *reverse*, consists in reversing of the order of distribution's multi-index from $D_{a_1 \ldots a_s}$ to $D_{a_s \ldots a_1}$. For example, $reverse(D_{1213}) = D_{3121}$. The second operation, termed *reorder*, restores regularity of the distribution's multi-index (i.e., after reverse operation). For example, $reorder(D_{3121}) = D_{1232}$. Both of these basic operations are linear in the order of s.

All distributions, being stored in an array in the lexical order, are indexed with numbers from 0 to $\varpi_s - 1$. The key for fast managing of the distributions is to find a method of very fast calculating of this index for a given distribution $D_{a_1 \ldots a_s}$. We will call this operation as *hash*. Our hashing function requires an auxiliary array `tab[x][y][z]`. Each entry of the `tab` array corresponds to the offset in the given lexical order. Three dimensions of this array represent triples (a_i, i, σ_i), where $\sigma_i = \max(a_1, \ldots, a_{i-1})$ and the entry of the `tab` is the index that the distribution $D_{a_1 \ldots a_s}$ would have if the index of $D_{a_1 \ldots a_{i-1} 1 \ldots 1}$ was 0. For example, the entry `tab[4][6][3]` represents the difference in positions in the lexical order between distributions $D_{a_1 \ldots a_5 1 \ldots 1}$ (where $\max(a_1, \ldots, a_5) = 3$) and $D_{a_1 \ldots a_5 4 1 \ldots 1}$.

The array `tab` can be easily filled during the creation of the distribution's representation if we modify Algorithm 4.1 into recursive Algorithm 4.2 with the initial call `fun(`D_1`,1)`

Having `tab` initialized by Algorithm 4.2, all what we need to do to obtain the index

Algorithm 4.2 Representation of the distributions in the lexical order – a recursive algorithm with filling of the `tab` array.

```
fun(D_{a_1...a_i}, σ_i)
begin
    k = 0
    for j = 1 to σ_i + 1
        tab[j][i + 1][σ_i] = k
        if i + 1 = s
            A ← D_{a_1...a_i j}
            k = k + 1
        else
            k = k + fun(D_{a_1...a_i j}, max(σ_i, j))
        endif
    endfor
    return k
end
```

of the distribution $D_{a_1 \ldots a_s}$ is to iterate over all indexes a_i, $1 \leq i \leq s$ and sum up the values of `tab` corresponding to these indexes. As we see, the hashing function is linear in the order of s.

Θ calculation

In order to obtain Θ from (4.6), we need to calculate all Θ_i, $0 \leq i < s$, where $i = 0$ stands for a case with no recombination and each $i > 0$ stands for a case with recombination after the ith locus. For a given recombination site, the probability transition matrix may be obtained based on (4.5) by analyzing all possible coalescent events in death/birth process. As we have already mentioned, these events are identified by the triples (j, k, m), $1 \leq j, k, m \leq 2N$, where j is the number of deceased individual replaced by the individual made as a result of recombination of individuals k and m. For a given distribution after coalescent event $\tilde{D}_i$, $0 \leq i < \varpi_s$, each triple (j, k, m) point at exactly one distribution D_j, $0 \leq j < \varpi_s$ leading to this $\tilde{D}_i$ as a result of the coalescent event. The probability of each triple is equal to $\frac{1}{(2N)^3}$. Thus, the algorithm to calculate Θ_i seems to be straightforward (Algorithm 4.3).

Algorithm 4.3 Calculation of the matrix Θ_i. The function `getDistr` is a linear function in the order of s that returns the distribution which, in a death/birth process described by parameters, leads to the distribution given as a parameter.

$A \leftarrow$ distributions in lexical order (indexed from 0 to $\varpi_s - 1$)
$\forall_{0 \leq x,y < \varpi_s} \Theta_i[x][y] = 0$
foreach $\tilde{D}_x \in A$
    foreach $(j,k,m), 1 \leq j,k,m \leq 2N$
        $D_y \leftarrow$ getDistr$(\tilde{D}_x, (j,k,m), i)$
        $I_x = \text{hash}(\tilde{D}_x)$
        $I_y = \text{hash}(D_y)$
        $\Theta_i[I_x][I_y] = \Theta_i[I_x][I_y] + \frac{1}{(2N)^3}$
    endfor
endfor

Unfortunately, in the Algorithm 4.3 we need to analyze all $(2N)^3$ triples (j,k,m). It causes two major problems: (i) depending on the value of population size makes it more difficult to obtain the symbolic Θ_i matrix (which is independent of the $2N$ value) and (ii) the algorithm is extremely time inefficient even for small populations. To overcome these difficulties we can notice that there exist groups of triples that lead to the same distribution's changes. To see this, let σ_x be the number of distinct characters in the multi-index corresponding to $\tilde{D}_x = \tilde{D}_{a_1 \ldots a_s}$. Then, $\sigma_x = \max(a_1, \ldots, a_s)$. Note that if $j > \sigma_x$, $\tilde{D}_x$ is D_x. Next, consider the recombination event described by (j,k,m) where $j \leq \sigma_x$, $k > \sigma_x$ and $k \neq m$, and assume that it leads from $\tilde{D}_x$ to D_y. Then, any recombination event (j,k',m) where $k' > \sigma_x$ and $k' \neq m$ also leads from $\tilde{D}_x$ to D_y. The same is true if roles of k and m are interchanged. Also if $j \leq \sigma_x$, $k = m > \sigma_x$ and $\tilde{D}_x$ leads to D_y, then $\tilde{D}_x$ leads to D_y for all other events described by (j,k',m') where $k' = m' > \sigma_x$. Hence, the triples (j,k,m) naturally divide into six classes and the computation of Θ_i may be performed in the following six steps:

1. the case where $j > \sigma_x$
2. the case where $1 \leq j,k,m \leq \sigma_x$
3. the case where $1 \leq j,k \leq \sigma_x$ and $m > \sigma_x$

Table 4.2: Division of the $2N \times 2N \times 2N$ space of (j,k,m) triples into six distinguish classes. Each triple (j,k,m) describes a single death/birth process (individuals k and m recombine and form a new individual that replaces deceased individual j). We apply the division of the space of the triples for each distribution $\tilde{D}_{a_1 \ldots a_s}$ and $\sigma = \max(a_1, \ldots, a_s)$. One can notice that the given probabilities can be represented as a linear combination of a^0, a^1, a^2 and a^3 where $a = \frac{1}{2N}$. Therefore, obtaining of the symbolic matrices Θ_i, $0 \leq i < s$ is straightforward by using slightly modified Algorithm 4.3.

Case number	Triple's requirements	Number of states	Probability of each state
1	$j > \sigma$	1	$1 - \frac{\sigma}{2N}$
2	$1 \leq j,k,m \leq \sigma$	σ^3	$\frac{1}{(2N)^3}$
3	$1 \leq j,k \leq \sigma, m > \sigma$	σ^2	$\frac{1}{(2N)^2} - \frac{\sigma}{(2N)^3}$
4	$1 \leq j,m \leq \sigma, k > \sigma$	σ^2	$\frac{1}{(2N)^2} - \frac{\sigma}{(2N)^3}$
5	$1 \leq j \leq \sigma, k = m > \sigma$	σ	$\frac{1}{(2N)^2} - \frac{\sigma}{(2N)^3}$
6	$1 \leq j \leq \sigma, k > \sigma, m > \sigma, k \neq m$	σ	$\frac{1}{2N} - \frac{2\sigma+1}{(2N)^2} + \frac{\sigma^2+\sigma}{(2N)^3}$

4. the case where $1 \leq j,m \leq \sigma_x$ and $k > \sigma_x$

5. the case where $1 \leq j \leq \sigma_x$, and $k = m > \sigma_x$

6. the case where $1 \leq j \leq \sigma_x, k,m > \sigma_x$ and $k \neq m$

The total number of triples that need to be analyzed in each of these steps and the probability values corresponding to each triple are presented in Table 4.2.

4.4.3 Time and memory complexity of the main program

The memory complexity $M(s)$ is a sum of space used by two matrixes used in the algorithm. The symbolic matrix is built of ϖ_s^2 triples of coefficients of 4 bytes numbers and the numerical matrix is obtained from the symbolic matrix by multiplying these coefficients by consecutive powers of $\frac{1}{2N}$ (and 1 is added on the main diagonal). Hence, $M(s) = M(\text{SymbolicMatrix}) + M(\text{NumericMatrix}) = 12\varpi_s^2 + 8\varpi_s^2 = 20\varpi_s^2[\text{Byte}]$, each number in the numerical matrix using 8 bytes. For example, $M(8) = 340\text{MB}$ and $M(9) = 8.5\text{GB}$.

We perform s^3 iterations (actually, σ^3 iterations) and we use s iterations to transform a multi-index involved into its regular form during the computation of each row

of the matrix. Hence, time-complexity of calculating each row is of the order of s^4. Taking into account the initialization process, we obtain that the time complexity is $\mathrm{O}(s^4\varpi_s + \varpi_s^2)$.

4.5 Results

4.5.1 Stationary distributions

We calculate the stationary distribution π by iteratively multiplying of the transition matrix Θ by itself until we reach the matrix with each row being almost equal to any other row of the matrix. The matrix after the kth iteration is equal to Θ^{2^k}. We assume that the two rows of the matrix are considered equal if all the differences of the values of the corresponding entries of both rows are lower than the chosen precision (usually equal to 10^{-6}).

The numerical calculations show that:

- as $r = \sum_{i=1}^{s-1} r_i$ increases and $2N$ is fixed, the role of π_1 in the stationary distribution decreases to 0, while that of π_{ϖ_s} increases to 1 (Figure 4.1).
- with the growth of r, each π_i where $1 < i < \varpi_s$ initially increases to a maximal value, and then decreases to zero (Figure 4.2).
- if $r_1 = r_2 = ... = r_{s-1}$, tuples of distributions related by symmetry, such as D_{11112} and D_{12222}, reach the maximal value at the same time (Figure 4.2).

This suggests that with the growth of r, the probability mass tends to concentrate close to π_{ϖ_s}. This intuition is supported by Figure 4.3 where the expected number of recombination events E_r leading to a distribution in the stationary state is shown to grow with r. The value is calculated as follow: $E_r = \sum_{i=1}^{\varpi_s} \pi_i \gamma_i$, where γ_i is the number of recombination events needed to obtain the ith distribution. In this case we assume that only one recombination event may appear after each locus. Then, for each distribution, the number of recombination events leading to it may be calculated as the number of the consecutive pairs of loci descended from the different individuals. For example, to obtain distribution D_{1223} exactly two recombination events are required (after the first and after the third locus).

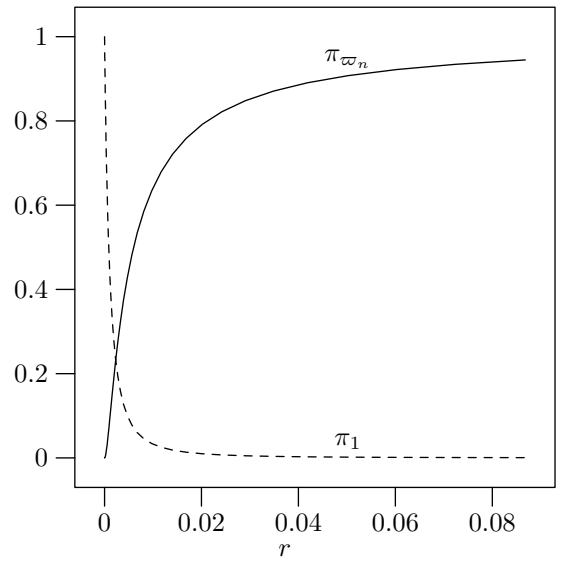

Figure 4.1: Values of the first (π_1) and the last (π_{ϖ_s}) entry of the stationary distribution for the model with five loci as a function of the recombination rate with constant population size $2N = 1000$.

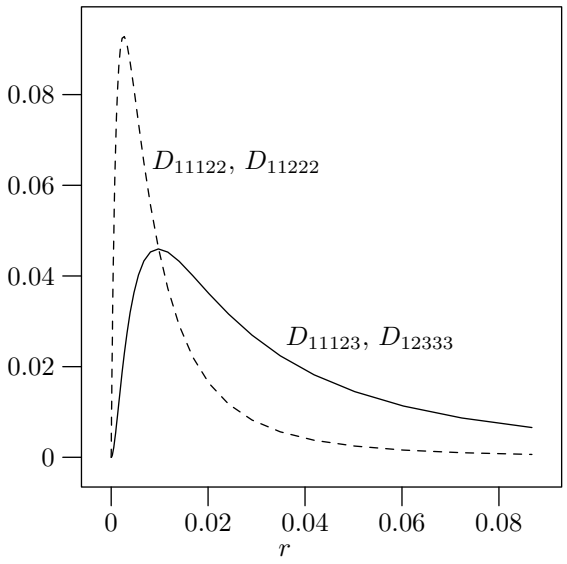

Figure 4.2: Examples of values of entries of the stationary distribution as a function of the recombination rate for constant population size $2N = 1000$. Since we assume $r_1 = r_2 = ... = r_{s-1}$ the entries for the distributions related by symmetry (such as D_{11123} and D_{12333}) are equal.

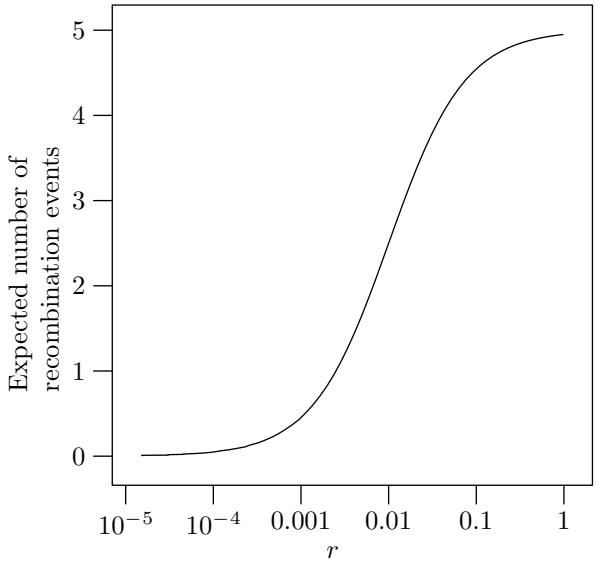

Figure 4.3: Expected number of recombination events for the model with six loci as a function of the recombination rate with constant population size $2N = 1000$.

However, the role of distributions close to the last one in the lexical order may decrease quite slowly (see Figure 4.4). Finally, the speed (based on the number of discrete generations) of reaching the stationary distribution of Θ is of the order of population size (Figure 4.5). We assume that the matrix reaches the stationary distribution when all its entries differ from the corresponding entries of the previously calculated stationary distribution by less than 10^{-6}. This agrees well with the fact that the time to the most recent common ancestor is of the order of population size.

4.5.2 Spectral gap

To calculate the spectral gap of the Θ we use separate computer program (see Section 4.4.1). The results obtained by us are intuitively clear: the speed of convergence (indicated by the lower values of spectral gap) decreases when the number of loci (Figure 4.6) or population size (Figure 4.7) increases.

4.5.3 Comparison with Wright-Fisher Hudson's model

Hudson's algorithm (90, 91, 93) is a well known standard coalescent approximation of the Wright-Fisher model. Wright-Fisher model assumes constant finite population size of $2N$ individuals and discrete generations with geometric distribution times between

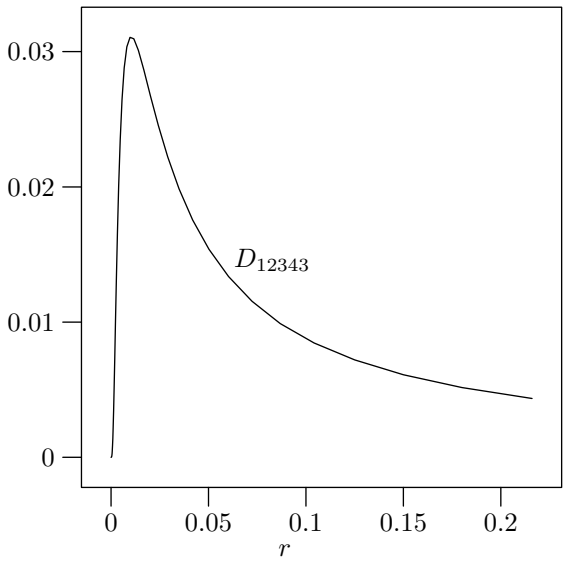

Figure 4.4: An entry D_{12343} of the stationary distribution close to the last distribution in lexical order for five loci as a function of the recombination rate for constant population size $2N = 1000$.

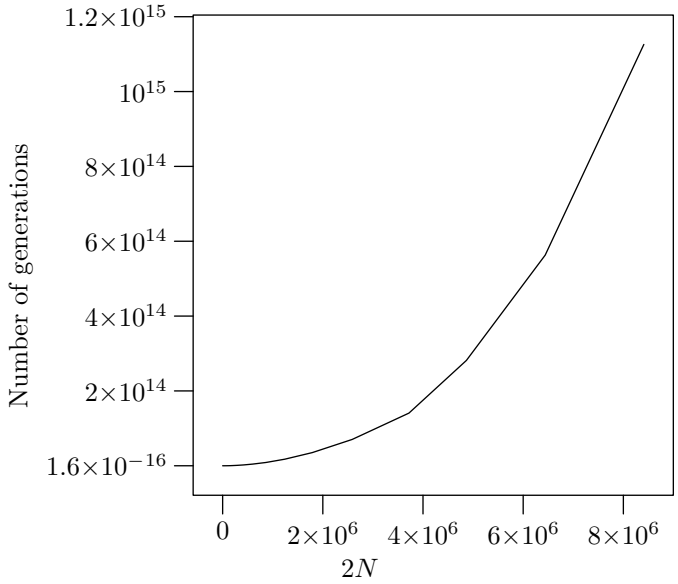

Figure 4.5: Number of discrete generations required for transition matrix Θ to reach, with a given precision, the stationary distribution in each row, as a function of the population size. We assume that the matrix reaches the stationary distribution when all its entries differ from the corresponding entries of the stationary distribution by less than 10^{-6}.

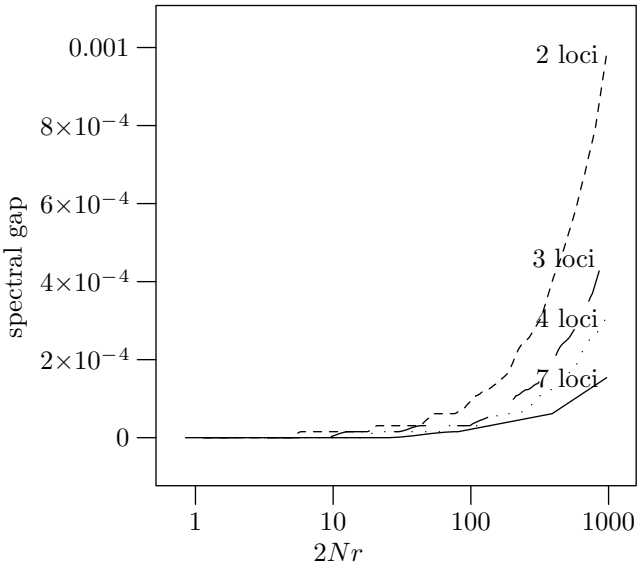

Figure 4.6: The spectral gap of Θ as a function of $2Nr$ coefficient calculated for models with different number of loci and constant population size $2N = 1000$.

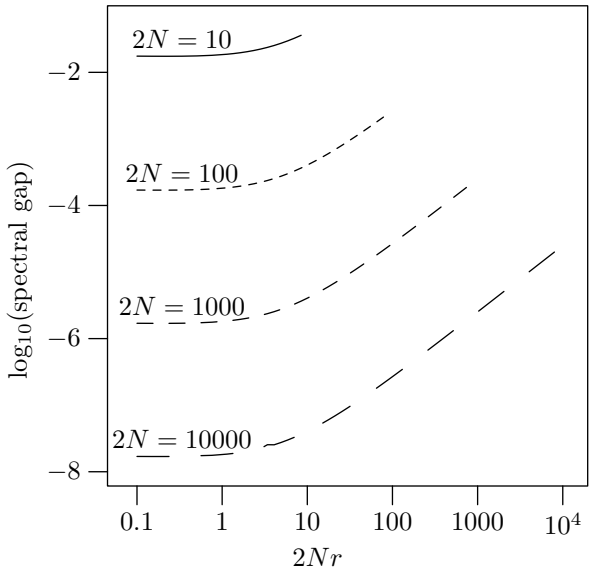

Figure 4.7: The spectral gap of Θ for the model with five loci as a function of $2Nr$, for various population sizes. Notice that the population size has a significant influence on the value of the spectral gap. Increasing the population size ten times results in decreasing the value of the spectral gap by about hundred times.

adjacent events. These assumptions are true for both, our and basic Hudson's model.

The difference is that our model, being a Moran model, incorporates a lifetime of each individual as an exponential random variable with parameter $\frac{2}{\lambda}$. Mutations in the Hudson's model are incorporated into genealogy and placed randomly on branches under infinite-sites model according to a Poisson process. Another model of mutations may be considered only by using sampled gametes as input to other model (48). Our model is not limited to any specific mutation model.

In both models a recombination event can occur between any two sites. The number of these sites is finite, constant and specified by the user. An important difference in the recombination model between both models lies in the fact that in the Hudson's model the recombination event is independent from the coalescent event.

Hudson (90), based on results obtained by Griffiths (66), gives the following formula for the covariance of the number of segregating sites at two loci in a sample of n gametes:

$$Cov(c_1, c_2) = 4N^2\mu^2 Cov(t_1, t_2) = 4\mu^2 Cor(t_1, t_2) \tag{4.15}$$

where c_i is the number of segregating sites at the ith locus, t_i is the time (in units of $4N$) to the MRCA of the entire sample at the ith locus, $R = 4Nr$ and

$$Cor(t_1, t_2) = f(R) = \frac{R + 18}{R^2 + 13R + 18}. \tag{4.16}$$

We conducted simulations using a modification of the Hudson's algorithm in order to estimate the correlation of the time to the MRCA at both loci in our model. Figure 4.8 presents results obtained for $n = 2$. The correlation in our model is higher than the correlation derived by Hudson for his model. That is intuitively expected since incorporating of the lifetimes of individuals provides dependence between individuals from the sample. We applied the mean square error interpolation to our model and received the following formula for the correlation:

$$Cor(t_1, t_2) = \frac{R + 32}{R^2 + 10R + 32}. \tag{4.17}$$

If we remove the information about times of birth of each individual from our model and scale the recombination rate by the times between two events (the times generated with an exponential distribution with parameter dependent on the sample size), then we obtain exactly the same formula for the correlation of the time to the MRCA at both loci as the one derived by Hudson (Figure 4.9).

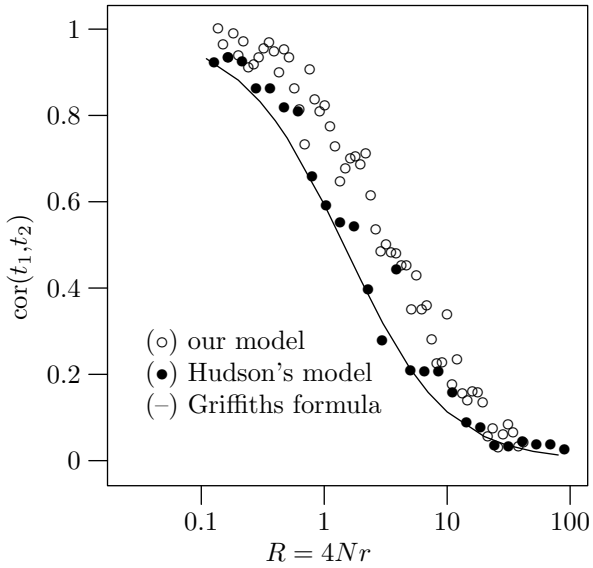

Figure 4.8: The correlation of the time to the MRCA at two loci in a sample of size 2 as a function of $R = 4Nr$. Results for both models, our and Hudson's, were obtained by Monte Carlo coalescent method under assumption of $n \ll 2N$. Results obtained by simulations for the Hudson's model are consistent with theoretical results provided by the Griffiths-Hudson formula (4.16). On the graph, expression (4.16) is depicted by a continuous line.

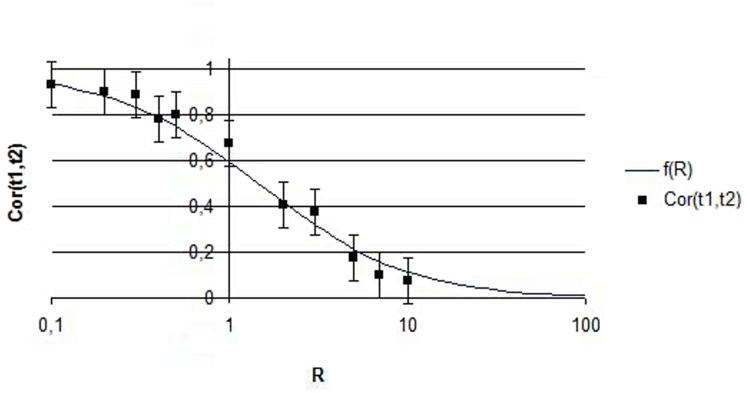

Figure 4.9: The correlation of the time to the MRCA at two loci in a sample of size 2 as a function of $R = 4Nr$. We removed from the model the information about times of death and scaled the recombination rate by the times between adjacent events. In that case the formula for the correlation in our model is similar to (4.16). The black squares are estimates of the correlation obtained by simulations. The continuous line depicts the relationship (4.16). Standard error intervals are indicated.

We obtained the results presented in Figure 4.8 under assumption that $n \ll 2N$. However, the correlation for small values of $2N$ ($2N < 6n$) slightly changes if we discard this assumption (Figure 4.10).

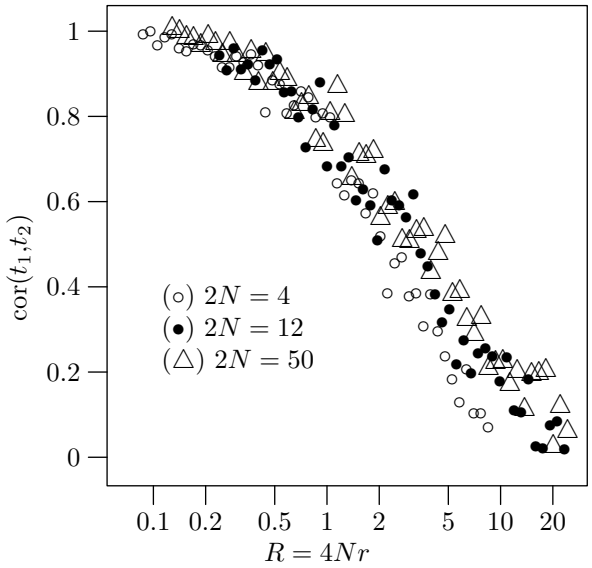

Figure 4.10: The correlation of the time to the MRCA at two loci in a sample of size 2 as a function of $R = 4Nr$ obtained by a standard Monte Carlo coalescent method. We can observe the influence of the $n \ll 2N$ assumption for a small values of $2N$ (equal to 4 in this case).

5

Genetic drift model in population with time-varying size

5.1 Preliminaries

We assume that the population has evolved from time $t = 1$ to the present time T. Population size at time t is denoted by N_t with $N_1 \geq 1$. The sequence $\{N_t\}_{t\in\{1,2,\ldots,T\}}$ is generally a discrete-time and discrete-state random process. To focus attention we assume that it is a Markov chain. We also assume multinomial sampling from a given generation's pool conditional on the number of individuals in the generation.

By $\tau_{i,t}$, $1 \leq \tau_{i,t} \leq T-1$ we denote the time (counted backwards) to the MRCA of a sample of i individuals living at time t. The distribution that we need to calculate is $\{P(\tau_{n,T} = t), t = 1, 2, \ldots, T-1\}$.

5.1.1 Wright-Fisher model with time-varying population size

We consider a coalescent tree built for a sample of size n randomly drawn from the current generation of the population evolved by the Wright-Fisher model. The tree contains $n-1$ coalescent events and we denote the random coalescence times of these events by $H_n, H_{n-1}, \ldots, H_2$ and their realizations by $h_n, h_{n-1}, \ldots, h_2$. All times (in generations) are counted backwards from the current generation at time 0. We assume that $h_2 \geq h_3 \geq \cdots \geq h_n \geq 0 = h_{n+1}$ and that $N(t)$ is the effective population size at time t. Then, according to (67), the joint probability density function of the times

$H_2, H_3, \ldots, H_n$ has the form:

$$p(h_2, h_3, \ldots, h_n) = \prod_{k=2}^{n} \frac{\binom{k}{2}}{N(h_k)} \exp\left(- \int_{h_{k+1}}^{h_k} \frac{\binom{k}{2}}{N(z)} \mathrm{d}z \right) \tag{5.1}$$

Equation (5.1) allows us to calculate the distribution $P(\tau_{n,T} = t)$ by averaging of the $p(h_2, h_3, \ldots, h_n)$ values over all possible realizations of the coalescent tree. Unfortunately, this method can be extremely time inefficient for larger n or complex $N(t)$ dependency. Nevertheless, a modification of formula (5.1) yields the exact value of the distribution $P(\tau_{n,T} = t)$ for a population with known time-varying size and for small n. Below we consider the case $n = 2$.

For consistency we assume that population evolved from time $t_1 = 1$ to the current generation at time T. The population, over its evolution, may experience discrete events leading to a change of the growth rate.We denote the number of these events by $m - 2 \geq 0$ and the times of their occurrence by $t_2, t_3, \ldots, t_{m-1}$, where $1 = t_1 < t_2 < \cdots < t_{m-1} < t_m = T$. The population size is equal to $N_i(t)$ for $t \in (t_i, t_{i+1})$. Functions $N_i(t)$ may be chosen from families such that expression (5.1) is expressed in the terms of elementary functions. As a result we obtain the following probability density function for the time to the MRCA of a pair of individuals drawn from the population at time T:

$$p(\tau_{2,T} = t) = \frac{1}{N_k(t)} \exp\left(- \int_{t}^{t_{k+1}} \frac{1}{N_k(z)} \mathrm{d}z \right) \prod_{i=k+1}^{m-1} \left(\exp\left(- \int_{t_i}^{t_{i+1}} \frac{1}{N_i(z)} \mathrm{d}z \right) \right), \tag{5.2}$$

where $t_k \leq t \leq t_{k+1}$.

One may notice that if we set $m = 2$ and $N_1(t) = N$, i.e., a constant size population, then formula (5.2) gives us the well-known exponential distribution of the time to the MRCA of a pair of chromosomes (181).

5.1.2 Bobrowski's formula for the distribution of the time to the MRCA

Let us consider the case of $n = 2$. Two individuals at generation $t + 1$ are descendants of the same member of generation t with probability $p_t = 1/N_t$ and with probability $q_t = 1 - p_t$ they are descendants of two different individuals. As derived in (18),

$$P(\tau_{2,T} = t) = \prod_{k=T-t}^{T-1} q_k - \prod_{k=T-t-1}^{T-1} q_k = p_{T-t-1} \prod_{k=T-t}^{T-1} q_k, \tag{5.3}$$

where for mathematical consistency we set $p_0 = 1$ and $q_0 = 0$.

Let $\sigma_{i,t}$ be the time to the first coalescence of the genealogical lineages of the sample of i individuals of generation t. If $\sigma_{n,T} = s$ and there are k distinct ancestors of the sample at time $T - s$, where $1 \leq s \leq T - 1$, then

$$\tau_{n,T} = \sigma_{n,T} + \tau_{k,T-s} \tag{5.4}$$

and the summands are independent conditionally on $\{N_t\}$.

The probability that m members of generation $t + 1$ have exactly k ancestors at generation t is ((17) p. 352):

$$q_{m,k,t} = \frac{S_{m,k}\binom{N_t}{k}k!}{N_t^m}, \tag{5.5}$$

where $S_{m,k} = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^m$ is the Stirling number of the second kind (63). Introducing $p_{m,t} = q_{m,m,t}$ for $m \leq N_t$, and $p_{m,t} = 0$ otherwise ($p_{m,t}$ is the probability of m members of generation $t + 1$ having m ancestors at generation t), we obtain

$$P(\sigma_{n,T} = s) = q_{n,k,T-s}\prod_{u=1}^{s-1} p_{n,T-u}. \tag{5.6}$$

Hence, summing over all k and s, we obtain based on (5.4),

$$P(\tau_{n,T} = t) = \sum_{s=1}^{T-1}\sum_{k=1}^{n-1} q_{n,k,T-s}\prod_{u=1}^{s-1} p_{n,T-u}P(\tau_{k,T-s} = t - s), \tag{5.7}$$

where $1 \leq t \leq T - 1$.

5.1.3 Time to the MRCA in a Galton-Watson process

Assume that one has the knowledge of the complete genealogical history of the population (full genealogical tree). Then, the simplest method to calculate the distribution of the time to the MRCA of a sample of size n is by tracing back the lineages of all possible samples of that size to their MRCA and using the frequency graph as an estimate. Unfortunately, this approach is extremely time-inefficient and cannot be used for larger populations or larger sizes of sample – the exact number of samples to trace is equal to $\binom{N_t}{n}$. In that case, the only method to estimate the distribution is to reduce the number of examined samples using Monte-Carlo simulations. We can obtain the results by averaging the times calculated for a fixed number of samples randomly drawn from

the population. A sample of individuals drawn from a population evolved according to the Galton-Watson process satisfies Lemma 5.1.

Lemma 5.1. $\tau_{a_1,a_2,\ldots,a_{i+1}} = max(\tau_{a_1,a_2,\ldots,a_i}, \tau_{a_1,a_{i+1}})$, *where* τ *is the time to the MRCA of a sample of individuals listed as indices.*

Proof. Let us denote $\tau_{a_1,a_2,\ldots,a_i} = \alpha$, $\tau_{a_1,a_{i+1}} = \beta$, η_1 is the MRCA of the sample $(a_1,\ldots,a_i)$ and η_2 is the MRCA of a pair (a_1, a_{i+1}). Following observations are based on the fact that a_1, η_1 and η_2 belong to the same lineage in the genealogical tree. If $\beta \leq \alpha$, then η_2 is a descendant of η_1 and $\tau_{a_j,a_{i+1}}$, for $1 \leq j \leq i$, cannot be greater than α. In that case $\tau_{a_1,a_2,\ldots,a_{i+1}} = \alpha$. Otherwise, $\tau_{\eta_1,a_{i+1}} = \beta$, and $\tau_{a_j,a_{i+1}} = \beta$ for each $1 \leq j \leq i$. Thus, $\tau_{a_1,a_2,\ldots,a_{i+1}} = \beta$. □

Lemma 5.1 limits to $n-1$ the number of calculations of the time to the MRCA of two individuals required to calculate the time to the MRCA of a sample of size n.

5.2 Model derivation

As we have already mentioned, the well-known formula (5.1) does not allow us to obtain the expected distribution of the time to the MRCA given by $P(\tau_{n,T} = t)$ for required values of n, T or N_i. The Bobrowski's formula (5.7) is also insufficient. Admittedly, it gives us a direct method to calculate this distribution but the recurrence in the formula and the use of the Stirling numbers make the calculation infeasible even for relatively small T, n and N_i (i.e., if $T \approx 100$ and $n = 3$, the calculation cannot be completed in a feasible amount of time). However, the computation may be reorganized in order to deal with the stated problem.

Let $\alpha_{t,k}$ be the probability that the sample has exactly k ancestors at time t. Obviously, $\alpha_{T,n} = 1$ and $\alpha_{T,i} = 0$, $i \neq n$. Hence, we get

$$\alpha_{t,k} = \sum_{i=k}^{n} \alpha_{t+1,i} q_{i,k,t} \tag{5.8}$$

where q values are given by (5.5).

Based on (5.8) we are able to calculate all α values if we know the values of the q probabilities. Moreover, the probability that the MRCA of the sample exists at time t

is equal to $(\alpha_{t,1} - \alpha_{t+1,1})$, the difference between the probabilities that the sample has one ancestor at times t and $t+1$. Hence,

$$P(\tau_{n,T} = t) = (\alpha_{T-t,1} - \alpha_{T-t+1,1}) \tag{5.9}$$

The $\alpha_{t,i}$ probabilities also define the distribution of the number of lineages ancestral to a sample at time t. These values may be used, for example, in the methods that estimate the whole population history based on the history of the sample(s) from that population (124).

Finally, we get the following two theorems, Theorem 5.1 and Theorem 5.2. In the proofs of these theorems we repeatedly use the two following recurrence relations: $S_{n,1} = S_{n,n} = 1$ and $S_{n,k} = S_{n-1,k-1} + kS_{n-1,k}$.

Theorem 5.1. *The probability values $q_{n,k,t}$ satisfy the following equations:*

$$q_{1,1,t} = 1, \qquad 1 \le t \le T \tag{5.10}$$

$$q_{i+1,i+1,t} = q_{i,i,t}\frac{N_t - i}{N_t}, \qquad 1 \le t \le T, 1 \le i < n \tag{5.11}$$

$$q_{i+1,k,t} = \frac{W_{i,k}}{N_t} q_{i,k,t}, \qquad 1 \le t \le T, 1 \le k \le i < n \tag{5.12}$$

where $W_{i,k} = \frac{S_{i+1,k}}{S_{i,k}}$.

Proof. Equation (5.10) is obvious from (5.5). Two other formulas may be obtained as follow:

$$q_{i+1,i+1,t} = \frac{S_{i+1,i+1}\binom{N_t}{i+1}(i+1)!}{N_t^{i+1}} = \frac{1\frac{N_t!}{(N_t-i-1)!i!(i+1)}(i+1)!}{N_t^i N_t} = \frac{S_{i,i}\frac{N_t!}{(N_t-i)!i!}i!(N_t-i)}{N_t^i N_t} = q_{i,i,t}\frac{N_t-i}{N_t}$$

$$q_{i+1,k,t} = \frac{S_{i+1,k}\binom{N_t}{k}k!}{N_t^{i+1}} = \frac{S_{i+1,k}}{S_{i,k}N_t}\frac{S_{i,k}\binom{N_t}{k}k!}{N_t^i} = \frac{W_{i,k}}{N_t}q_{i,k,t}, \text{ where } W_{i,k} = \frac{S_{i+1,k}}{S_{i,k}}. \qquad \square$$

In Theorem 5.1 we introduced the $W_{n,k}$ sequences. The values of these sequences are defined by Theorem 5.2.

Theorem 5.2. *The values of $W_{n,k}$ satisfy the following equations:*

$$W_{i,1} = 1, \qquad 1 \le i \le n \tag{5.13}$$

$$W_{i,i} = W_{i-1,i-1} + i, \qquad 2 \le i \le n \tag{5.14}$$

$$W_{i,k} = k + W_{k-1,k-1} \prod_{j=k}^{i-1} \frac{W_{j,k-1}}{W_{j,k}}, \qquad 2 \leq k < i \leq n. \tag{5.15}$$

Proof. $W_{i,1} = \frac{S_{i+1,1}}{S_{i,1}} = 1$

$W_{i,i} = \frac{S_{i+1,i}}{S_{i,i}} = S_{i,i-1} + iS_{i,i} = \frac{S_{i,i-1}}{S_{i-1,i-1}} + i = W_{i-1,i-1} + i$

$W_{i,k} = \frac{S_{i+1,k}}{S_{i,k}} = \frac{S_{i,k-1}+kS_{i,k}}{S_{i,k}} = k + \frac{S_{i,k-1}}{S_{i,k}} = k + \frac{\frac{S_{i,k-1}}{S_{i-1,k-1}} \frac{S_{i-1,k-1}}{S_{i-2,k-1}} \cdots \frac{S_{k,k-1}}{S_{k-1,k-1}} S_{k-1,k-1}}{\frac{S_{i,k}}{S_{i-1,k}} \frac{S_{i-1,k}}{S_{i-2,k}} \cdots \frac{S_{k+1,k}}{S_{k,k}} S_{k,k}} =$
$k + \frac{W_{i-1,k-1} W_{i-2,k-1} \ldots W_{k,k-1} W_{k-1,k-1}}{W_{i-1,k} W_{i-2,k} \ldots W_{k,k}} = k + W_{k-1,k-1} \prod_{j=k}^{i-1} \frac{W_{j,k-1}}{W_{j,k}}$ □

One may notice that $W_{n,k} = \frac{S_{n+1,k}}{S_{n,k}} = k + \frac{S_{n,k-1}}{S_{n,k}}$. From $\frac{S_{n,k-1}}{S_{n,k}}$ strictly decreasing with $n \to \infty$ (21) and based on (5.14) we can obtain that $W_{n,k}$ values are relatively small satisfying the following estimate: $W_{n,k} < n^2$.

5.3 Model implementation

The main algorithm uses Theorem 5.1 and Theorem 5.2 to calculate the distribution $P(\tau_{n,T} = t)$ in an efficient way. Besides this algorithm, we developed a framework that allows us to realize studies of the different scenarios of the population growth or mating schemas with a particular respect to the Galton-Watson process.

5.3.1 Main algorithm

The main algorithm uses the dynamic programming method (15, 29, 111) multiple times. The dynamic programming is a mathematical and computer algorithmic scheme for solving optimization problems. The method builds the final solution by expanding initial conditions (usually being the solutions for trivial cases) step by step into more complex cases based on a given formula. Time-efficiency of the method is determined by the number of states (equal to nT in our case) and the complexity of the final formula and is greatly improved by memorization of the partial solutions and their use in the succeeding steps. Thus, the dynamic programming is a perfect method to solve recurrence equations with given initial conditions and a simple recurrence expression.

Formulas (5.8) and (5.9) give us a direct expression that can be used to calculate the expected distribution. Algorithm 5.1 presents the dynamic programming implementation that realizes this task. The Q array used in Algorithm 5.1 contains the q probabilities given by (5.5).

Algorithm 5.1 Calculation of the distribution $P(\tau_{n,T} = t)$. The array $A[t][k]$, $1 \leq t \leq T, 1 \leq k \leq n$ stores the values of $\alpha_{t,k}$ defined by (5.8). The function `calcQ(t)` calculates the q probabilities at time t and stores them in the array $Q[m][k]$, $1 \leq k \leq m \leq n$.

$\forall_{1 \leq t \leq T, 1 \leq k \leq n} A[t][k] = 0$
$A[T][n] = 1.0$
for $t = T - 1$ downto 1
    calcQ(t)
    for $k = 1$ to n
        for $i = k$ to n
            $A[t][k] = A[t][k] + A[t+1][i] \cdot Q[i][k]$
        endfor
    endfor
endfor
for $t = T - 1$ downto 1
$P(\tau_{n,T} = t) = A[t][1] - A[t+1][1]$
endfor

To calculate all q probabilities, first we use dynamic programming to calculate all $q_{i,i,t}$ probabilities based on the recursive formula given by Theorem 5.1. Subsequently we extend these calculation to all $q_{n,k,t}$. Algorithm 5.2 presents the implementation of the calculation of the Q array used in Algorithm 5.1. By introducing the $W_{n,k}$ expression we avoid in our calculations the necessity of computing the Stirling numbers $S_{n,k}$, which are very large even for small n and k. To calculate the $W_{n,k}$ values we use Theorem 5.2. Algorithm 5.3 presents details of these calculations.

The total time complexity of the algorithm is of the order of $O(n^3 + n^2 T)$, where T is the number of discrete generations and n is the sample size. Thus, one may obtain the results in a short time even for $n \approx 10^3$ and the time period comparable to the time-span of modern humanity.

Algorithm 5.2 Calculation of the q probabilities at time t. N_t is the population size at time t.

```
calcQ(t)
begin
    ∀_{1≤i,j≤n} Q[i][j] = 0
    Q[1][1] = 1.0
    for i = 2 to n
        Q[i][i] = ((N_t − (i−1)) / N_t) Q[i−1][i−1]
    endfor
    for k = 1 to n
        for m = k + 1 to n
            Q[m][k] = (W[m−1][k] / N_t) Q[m−1][k]
        endfor
    endfor
end
```

Algorithm 5.3 Calculation of the $W_{n,k}$ values.

```
∀_{1≤i,j≤n} W[i][j] = 0
for i = 1 to n
    W[i][1] = 1.0
endfor
for i = 2 to n
    W[i][i] = W[i−1][i−1] + 1
    for j = 2 to i − 1
        a = W[j−1][j−1]
        for k = j to i − 1
            a = (W[k][j−1] / W[k][j]) a
        endfor
        W[i][j] = j + a
    endfor
endfor
```

5.3.2 Framework structure

Our algorithm works well for any evolutionary scenario that assumes a multinomial mating scheme. Thus, it may be possible to use our approach to examine different population models and compare results obtained by our method to other methods, usually based on simulations. To make conducting such experiments possible, we developed a framework that realize this task. The key feature of the program lies in the fact that we deliver to the person using our framework a set of ready-to-use functions including implementation of a few different generators of the pseudo-random values, calculation of several evolutionary parameters along with the implementation of the algorithm described in the previous section and sample implementation of a single genealogy. The only thing that is required from the user is to write a code (based on a given sample) that realizes the experiments. The user may also add his own mating scheme by deriving from Genealogy class. We attach a few examples describing of how to use our program. The framework is written in the C++ programming language and consists of four modules:

1. Module Distr contains implementation of the most commonly used distributions. To generate a pseudo-random value we use the MZT generator (123) with the period equal to $2^{144} \approx 10^{43}$.

2. Module Genealogy contains implementation of a base class describing a single genealogy tree along with a very efficient implementation of the model of population evolved according to the Galton-Watson process (see Section 5.3.2 for details).

3. Module Stats contains a set of functions calculating the values of a few evolutionary parameters. The most important part of this module contains an implementation of our algorithm calculating the distribution of the time to the MRCA.

4. Module GW performs experiments. Input genealogy data may be either delivered by user or obtained during the experiment by using Genealogy module.

Galton-Watson process implementation

In order to compare our method with simulation approach we applied our algorithm to the Galton-Watson process. We need the knowledge of the full generation-to-generation

sampling scheme to calculate the time to the MRCA of a sample drawn from the population evolved according to such a process. Unfortunately, for large populations, it requires enormous amount of data to be stored and analyzed, limiting the possible number of generations one can model in a feasible time. To solve this problem, we reduce the description of the genealogy by finding the "reduced" genealogy, limited to the individuals who left descendants at T. Only information about individuals from the last generation is stored. Each individual contains the specified additional data that allows to track back its genealogy.

All individuals in the population and lineages in the genealogy are identified in each generation by consecutive integer numbers starting from 1. Consecutive numbers are assigned to the individuals being the offspring of the same individual from the previous generation. Additionally, for each individual from the current generation with index x, we store the time z when its genealogical lineage originated and lineage y from which it evolved. Thus, each individual is described by the triple $(X, Y, Z) = (x, y, z)$, where x is unique for each individual from a given generation. We assume that if the individual with $X = i$ at generation $t \geq y$ has more than one descendant at generation $t+1$, the descendant identified by the lowest number (say, i') inherits all its genealogical information, whereas other descendants (identified by $i'+1$, $i'+2$, and so on) are marked as the individuals that originated in generation $t+1$ from the lineage i'. If, for example, an individual (x, y, z) of generation t has exactly three descendants at generation $t+1$, then these offspring will be identified by the following triples: (x', y, z), $(x'+1, t+1, x')$ and $(x'+2, t+1, x')$. The only individual at the first generation is described by the $(1, 1, 1)$ triple and each first individual in the population at any generation is identified by exactly the same triple $(1, 1, 1)$. Figure 5.1 shows a sample genealogy tree built in that way. Notice that the y value, corresponding to the direct ancestor of the individual, indicates the lineage index from the current generation. As an example, the 5th individual of the 4th generation in the figure evolved from the 5th individual of the 3rd generation but its $y = 4$ due to the fact that the first descendant of the ancestor has an index 4 in the 4th generation. This relation between x and y values of individuals from the same generations allows us to reconstruct the full genealogy of the Galton-Watson process of all non-extinct lineages.

As we can notice on Figure 5.1, assigning of a proper triple to the newly formed individual in the generation-to-generation scheme may not be straightforward in the

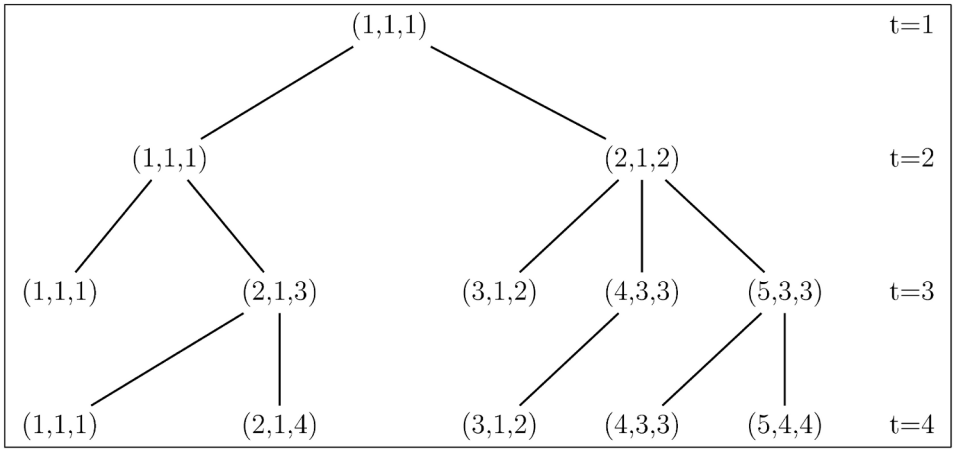

Figure 5.1: Sample genealogy tree evolved according to the Galton-Watson process. The entries of a triple describing each individual correspond to: the index of a lineage (an individual), the index of the ancestor and time when the individual originated.

case when its ancestral lineage extinct. Algorithm 5.4 demonstrates how these triples are obtained by using of a specialized map of lineages.

To find the time to the MRCA of two individuals a and b identified at time t by the triples (i_a, x_a, y_a) and (i_b, x_b, y_b) we need to perform an Algorithm 5.5. The algorithm is very fast with pessimistic time complexity equal to $O(\mathrm{N_T} + \mathrm{T})$ and logarithmic average time complexity. To find the time to the MRCA of a sample of size n we can use Lemma 5.1 that allows us to obtain this time by executing Algorithm 5.5 $n - 1$ times.

5.4 Results

5.4.1 Time to the MRCA of a sample drawn from a population experiencing a bottleneck event

As an example of the known time-varying size population we consider a population with a single bottleneck event. Our population history model assumes a long term constant population with the population size equal to N_1 followed by a reduction (bottleneck) of the population size to N_b ($N_b \leq N_1$) at generation T_b. Further on, the population grows exponentially in size to the final (current) generation T_f reaching a size equal to N_f. In the so-called hourglass scenario the reduction occurring in generation T_b is significant, whereas at the so-called longneck scenario $N_1 \approx N_b$. Figures 5.2, 5.3 and 5.4 compares the cumulative distributions of the time to MRCA for three different sizes of sample.

Algorithm 5.4 Creation of a new generation of individuals in the Galton-Watson process. Each individual a, $1 \leq a \leq N_t$ at time t is described by a triple (a, x_a, y_a). We denote a set of all individuals from the population at time t by A_t. Each individual a from the current generation has exactly w_a descendants at the following generation. We use the map L (being an array of size N_t) to map old lineages into new ones, the entry of L with a value of -1 stands for an extinct lineage.

```
k = 0
foreach a ∈ A_t
    if (w_a = 0)
        L[a] = −1
    else
        k = k + 1
        L[a] = k
        z = (z, x_z, y_z) ← a = (a, x_a, y_a)
        while (L[y_z] = −1)
            L[y_z] = k
            if (y_z = 0) //new lineage with the index 1 is formed
                x_z = 1
                break
            endif
            z ← (y_z, x_{y_z}, y_{y_z})
        endwhile
        y_z = L[y_z]
        A_{t+1} ← z
        for j = 2 to w_a
            A_{t+1} ← (k + j − 1, t, k)
        endfor
        k = k + w_a − 1
    endif
endfor
```

Algorithm 5.5 Calculation of the time to the MRCA of two individuals a and b in the Galton-Watson process. We assume that $i_a < i_b$.

```
findMRCA(a = (i_a, x_a, y_a), b = (i_b, x_b, y_b))
begin
    τ = x_b
    while (i_a ≠ y_b)
        τ = min(τ, x_b)
        b ← y_b = (i_{y_b}, x_{y_b}, y_{y_b})
        if (i_a > i_b)
            swap(a, b)
        endif
    endwhile
    return τ
end
```

All of presented populations starts 20000 generations in the past at the time $T_1 = 1$ with constant population size equal to $N_1 = 10000$. At the $T_b = 10000$th generation two of these populations experience a bottleneck event followed by exponential growth. Bottleneck event that appears at the T_bth generation reduces the size of populations to $N_b = 1000$ and $N_b = 9000$ for hourglass and longneck scenarios, respectively. For both of these populations the current population size is equal to $N_f = N_{20000} = 35000$. The results presented in Figure 5.2 (a case with $n = 2$) agree with its theoretical prediction calculated from the continuous-time approach based on expression (5.2).

5.4.2 Time to the MRCA of real populations

We apply our method to two real populations in order to calculate their distributions of the time to the MRCA. Our method requires knowledge of the values of the population size over the whole examined period. To meet that requirement, we assume that the modern census population sizes of the World and of Poland are as presented in Figures 5.5 and 5.6. Appendix A includes the references that we have used to obtain these estimates. The sources contain the World population size from the year 12000 BP along with a single older entry at the year 10^6 BP and the population size of Poland in the 1000 most recent years. One should be aware that the listed values, especially

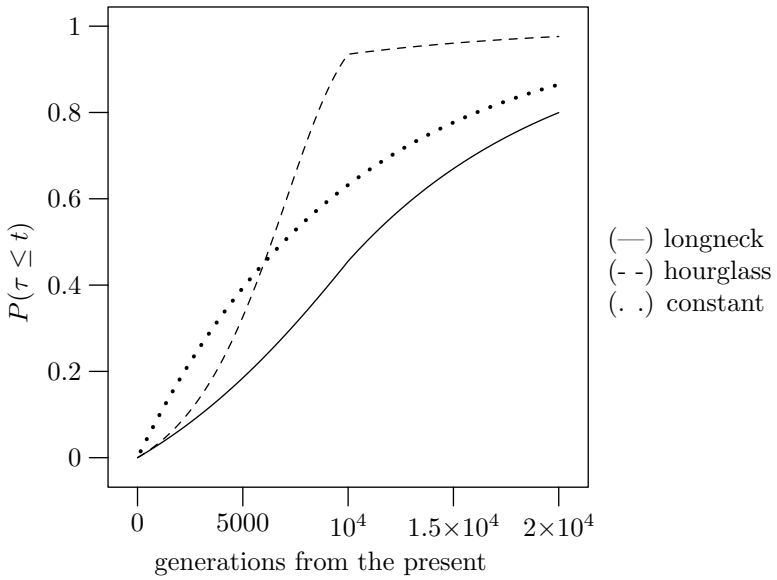

Figure 5.2: Time to the MRCA of a sample of size $n = 2$ in populations experiencing a bottleneck event. Three demographic scenarios are presented: longneck and hourglass bottleneck, and constant population size. In both populations experiencing the bottleneck $T_f = 20000$, $N_f = 35000$ and $T_b = 10000$. In the longneck scenario the population size decreases from $N_{9999} = N_1 = 10000$ to $N_{10000} = 9000$. In the hourglass scenario we assume $N_{9999} = 10000$ and $N_{10000} = 1000$. In the constant size population scenario $N = 10000$.

regarding the population size of the World B.C. and of Poland before the 1850s (when the censuses began), may not be correct. For example, the estimates of the population size of Poland in the late medieval centuries widely vary in different sources (even by 100%). The sizes at the prior times have been estimated using the growth rates listed by Kremer (113). We assume the World population growth rate to be equal to $3 \cdot 10^{-6}$ per year before the year 10^6 BP and the growth rate of Poland in the years before 1000 to be approximately proportional to the World's growth. In order to calculate the missing values of the population size, we divide (based on Figures 5.5 and 5.6 and using Kremer's ancient growth rates) the history of the World and of Poland into 8 and 4 time periods, respectively. In the case of Poland we assume years -10000, 1000 and 1850 as the boundaries. The boundary years for the World are as follows: -10000, -4000, -1000, 200, 1100, 1400 and 1950. We assume that the population size in each of these periods changes according to an exponential function. A single human generation in our model is assumed to last 25 years. The three most recent generations, assumed

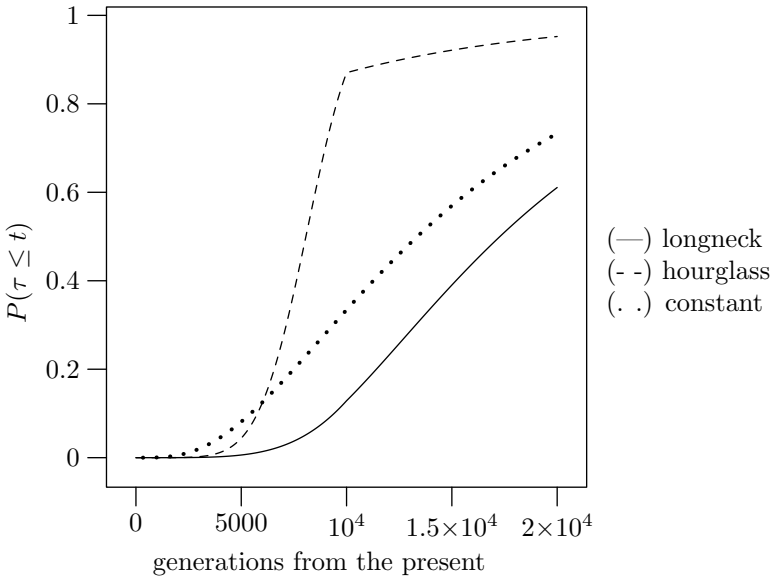

Figure 5.3: Time to the MRCA of a sample of size $n = 5$ in populations experiencing a bottleneck event. Three demographic scenarios are presented: longneck and hourglass bottleneck, and constant population size. In both populations experiencing the bottleneck $T_f = 20000$, $N_f = 35000$ and $T_b = 10000$. In the longneck scenario the population size decreases from $N_{9999} = N_1 = 10000$ to $N_{10000} = 9000$. In the hourglass scenario we assume $N_{9999} = 10000$ and $N_{10000} = 1000$. In the constant size population scenario $N = 10000$.

to have originated in years 1950, 1975 and 2000, are based on exact census data. The ratio of the human effective to census population size is usually estimated between 0.3 (101) and 0.5 (140). Taking into consideration the difference in the value of the effective population size between panmictic haploid population and the real human population (57) we assume the ratio to be equal to 0.25.

Figure 5.7 shows the result obtained for the population of Poland for recent 10^5 generations and Figure 5.8 depicts a similar result for the World population. The model does not take into consideration recombination events. Therefore, the obtained estimates of the expected time to the MRCA ($1.5 \cdot 10^6$ years for the World population and $1.5 \cdot 10^5$ years for the population of Poland) may apply only to non-recombining fragments of the genome (scale of single genes or below).

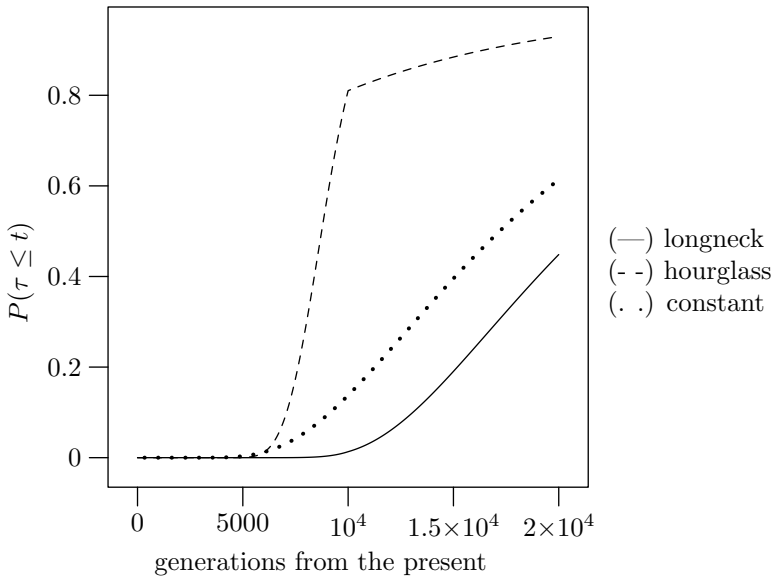

Figure 5.4: Time to the MRCA of a sample of size $n = 2$ in populations experiencing a bottleneck event. Three demographic scenarios are presented: longneck and hourglass bottleneck, and constant population size. In both populations experiencing the bottleneck $T_f = 20000$, $N_f = 35000$ and $T_b = 10000$. In the longneck scenario the population size decreases from $N_{9999} = N_1 = 10000$ to $N_{10000} = 9000$. In the hourglass scenario we assume $N_{9999} = 10000$ and $N_{10000} = 1000$. In the constant size population scenario $N = 10000$.

5.4.3 Time to the MRCA of the Galton-Watson population and comparison with direct simulation of the population process

We used our framework to generate a set of populations that have evolved according to the standard Galton-Watson process. In Figure 5.9 we present the cumulative distributions of the time to the MRCA averaged over 5000 non-extinct populations simulated over 100 generations. We arbitrarily assume that for each individual the number of its offspring is a random variable with a Poisson distribution with the parameter $\psi = 1.1$. In Figures 5.10, 5.11 and 5.12 we show how the time to the MRCA varies if we change the value of ψ (starting from Figure 5.10 we set the following values of ψ: 0.95, 1.0, and 1.1); we consider the case of $n = 3$. The gray area indicates the 95% confidence interval. Finally, in Figures 5.12 and 5.13 we compare two methods applied to the data used in Figure 5.9: (i) our method calculating the exact expected distribution of the time to the MRCA based on the population size history (Figure 5.12) and (ii) a Monte-Carlo

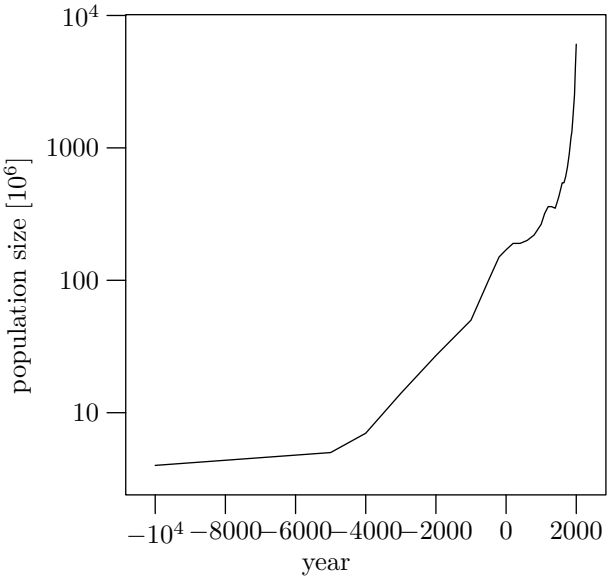

Figure 5.5: Demography of the World.

simulation over complete genealogies (Figure 5.13). In the Monte-Carlo simulation we calculate, for each of the 5000 genealogies, the time to the MRCA of 10^5 randomly drawn samples of size $n = 3$. Hence, the Monte-Carlo simulations, unlike our method, take into account the full genealogical history of the population. The mean values of the distribution obtained by both methods show almost complete agreement. Our method tends to decrease the variance. However, the results obtained by our method for a single genealogy may substantially differ from the real values due to the great variety of the possible genealogical histories observed for the Galton-Watson processes.

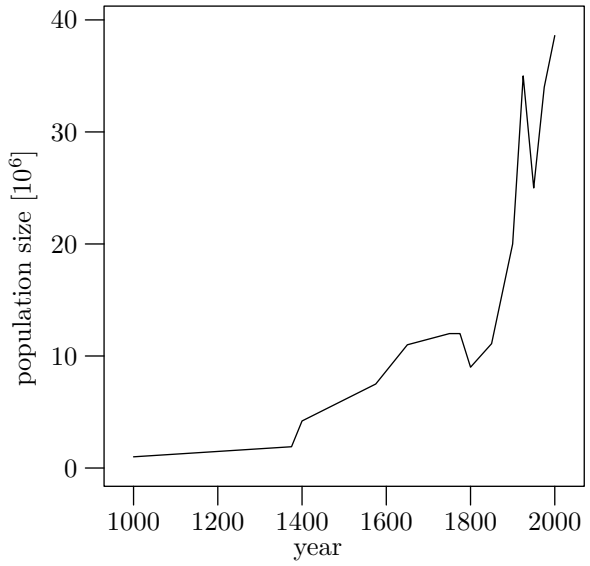

Figure 5.6: Demography of Poland.

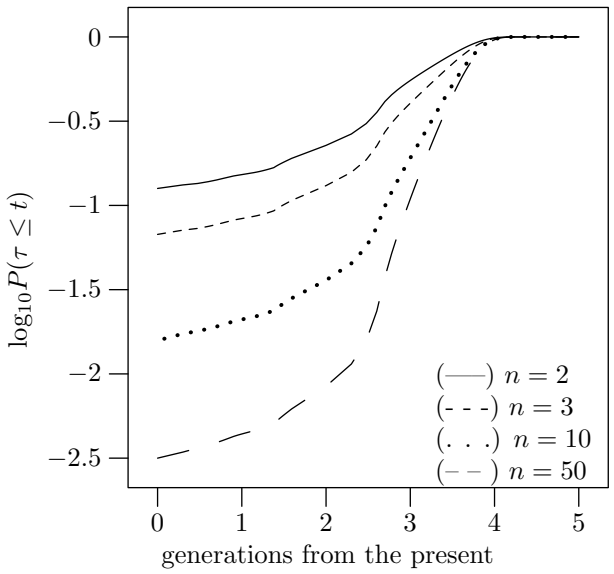

Figure 5.7: Time to the MRCA of the population of Poland for different sizes of a sample. The figure depicts the results for a period of recent 10^5 generations (counted backwards). We assume that each generation lasts 25 years.

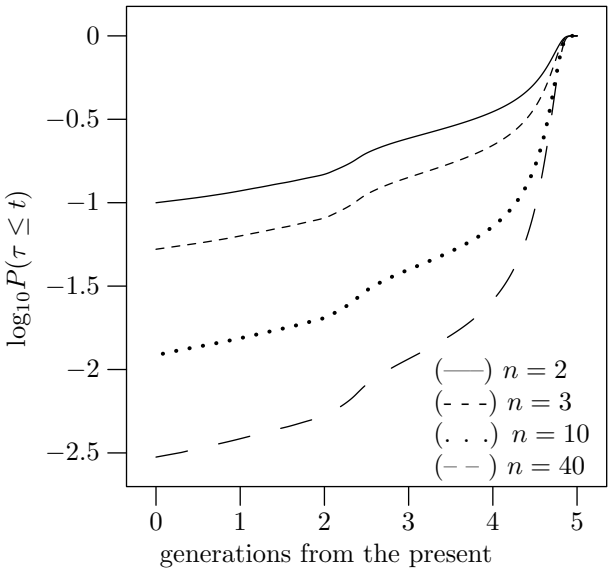

Figure 5.8: Time to the MRCA of the World population for different sizes of a sample. The figure depicts the results for a period of recent 10^5 generations (counted backwards). We assume that each generation lasts 25 years.

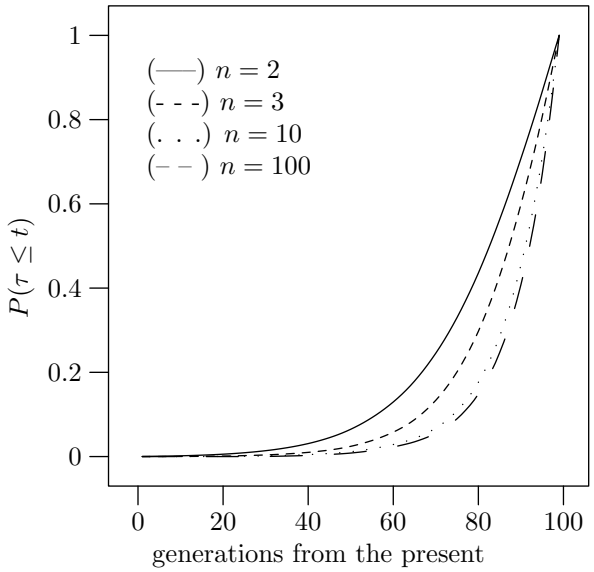

Figure 5.9: Cumulative time to the MRCA of a sample of size n from the population evolved over 100 generations according to the Galton-Watson process. The figure presents results for four different values of n. The results were averaged over 5000 non-extinct genealogies. We assume the Poisson offspring distribution parameter equal to 1.1.

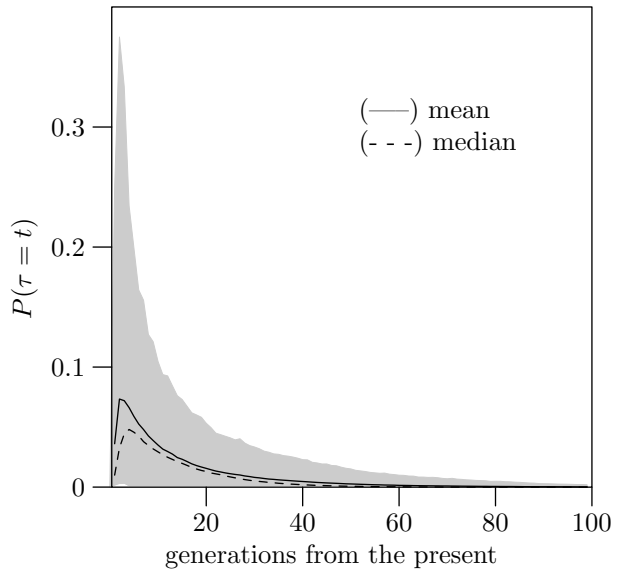

Figure 5.10: Time to the MRCA of a population evolved under the subcritical Galton-Watson process. The results were averaged over 5000 non-extinct genealogies evolved over 100 generations. $n = 3$ and $\psi = 0.95$. The gray area indicates the 95% confidence band.

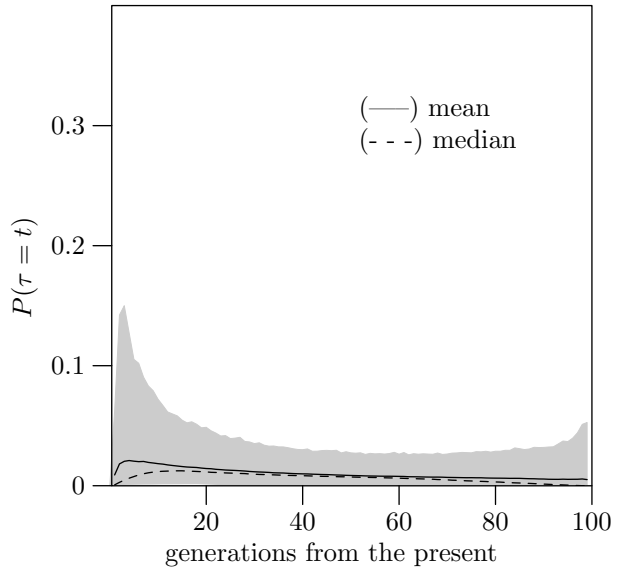

Figure 5.11: Time to the MRCA of a population evolved under the critical Galton-Watson process. The results were averaged over 5000 non-extinct genealogies evolved over 100 generations. $n = 3$ and $\psi = 1.0$. The gray area indicates the 95% confidence band.

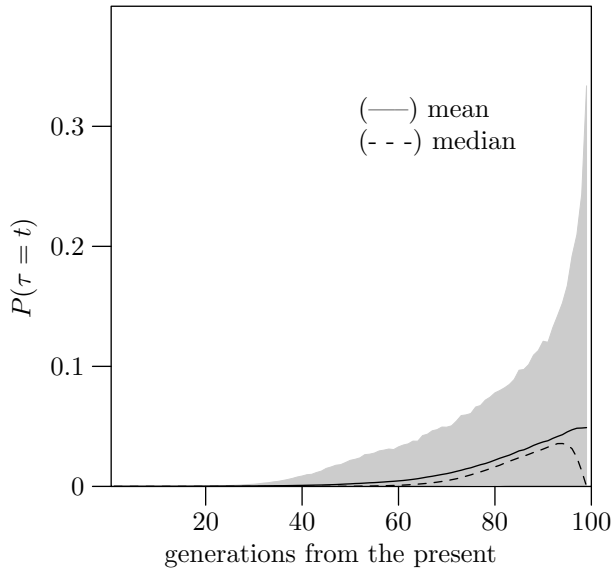

Figure 5.12: Time to the MRCA of a population evolved under the supercritical Galton-Watson process. The results were averaged over 5000 non-extinct genealogies evolved over 100 generations. $n = 3$ and $\psi = 1.1$. The gray area indicates the 95% confidence band.

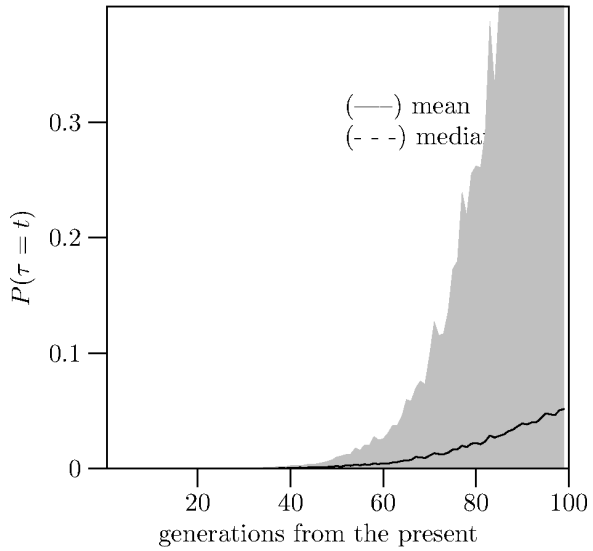

Figure 5.13: Comparison with a Monte-Carlo method. The figure presents results obtained by Monte-Carlo simulations for the case with $n = 3$ and $\psi = 1.1$. We determine the time to the MRCA for each population by calculating the time to the MRCA many times for different samples drawn from the last generation and averaging the results. The gray area indicates the 95% confidence band.

6

Demographic network model

6.1 Demographic network with merges, splits and migrations between populations

6.1.1 Description of the network

We consider a demographic network of populations evolving from a single ancestral population. The evolution of the network begins at time $t_0 = 0$ and continues forward in time. The network experiences three types of discrete events: merges of two populations into one, splits of a single population into two populations and migrations between any (possibly all) populations in the network. These events are chronologically ordered and occur at times t_i, $1 \leq i \leq I$ and $t_i \leq t_{i+1}$, where t_I is the present time. We allow more than one event to occur at the same time, but these events are distinguished from each other and are considered separately one after another according to a given order.

We denote the number of populations in the network in the time interval $[t_i, t_{i+1})$ as $\kappa_i \geq 1$, where $\kappa_0 = 1$ and $\kappa_i = \kappa_{i-1} + \gamma_i$ and γ_i is an indicator of change of the number of populations corresponding to the type of event occurred at time t_i. Depending on the type of event at time t_i, γ_i is equal to:

$$\gamma_i = \begin{cases} -1 & \text{for split} \\ 0 & \text{for migration} \\ 1 & \text{for merge} \end{cases} \tag{6.1}$$

Each population in the network is identified in the time interval $[t_i, t_{i+1})$ by a single index $k \in 0, 1, \ldots, \kappa_{i-1}$. The index of the population may change between two

consecutive time intervals. If the index of the population is k, $0 \leq k < \kappa_i$, in the time interval $[t_i, t_{i+1})$, then we denote it k', $0 \leq k' < \kappa_{i-1}$, in the previous time interval $[t_{i-1}, t_i)$. If at the time t_i population x splits, then

$$k = \begin{cases} k' & k' \leq x \\ x+1 & \text{for a newly created population} \\ k'+1 & k' > x. \end{cases} \tag{6.2}$$

If at the time t_i two populations with indices x and y $(x < y)$ merge, then

$$k = \begin{cases} k' & k' < y \\ k'-1 & k' > y, \end{cases} \tag{6.3}$$

the merged population has an index $k = x$ and the population with index $k' = y$ is removed from the network.

Migration event that occurred at time t is described by the matrix $M(t) = \{m_{xy}(t)\}$, $0 \leq x, y < \kappa_i$ with each entry m_{xy}, $0 \leq m_{xy} \leq 1$, $m_{xx} = 0$ equal to the migration rate from population x to y. Migration does not change indices of the populations.

The population size of population k at time $t \in [t_i, t_{i+1})$ is a function of time $N_{ik}(t)$ $0 \leq k < \kappa_i$.

6.1.2 Relations between populations in the network

We assume that we are given a demographic network described in the previous section. We consider a genetic feature associated with a haploid chromosome which can be sampled from any population existing in the network. We can describe this feature using an allelic space $\mathbb{A}$ containing $N_{\mathbb{A}}$ allelic types indexed from 1 to $N_{\mathbb{A}}$. We want to find an answer to the following question: What is the probability that a chromosome randomly sampled from population a at time t has the genetic feature of type j and that another individual from population b ($a = b$ is admissible) has the feature of type k ? These probabilities are entries of the joint distribution matrices $R_{ab}(t) = \{r_{ab}[j,k](t)\}$, $t \in [t_i, t_{i+1}), 0 \leq a, b < \kappa_i$ and $j, k \in \mathbb{A}$.

As above, if the index of the population is k, $0 \leq k < \kappa_i$, in the time interval $[t_i, t_{i+1})$, then we denote it k', $0 \leq k' < \kappa_{i-1}$, in the previous time interval $[t_{i-1}, t_i)$. Thus, the matrices $R_{ab}(t_i)$ and $R_{a'b'}(t_i - 0)$ indicate the joint distributions between two populations immediately after and immediately before the event occurred at time t_i, respectively.

If a split event occurs at time t_i, the allele on the chromosome in the splitting population is inherited by both progeny populations. Hence, we obtain the following identity for this case:

$$R_{ab}(t_i) = R_{a'b'}(t_i - 0) \tag{6.4}$$

If the event that occurs at time t_i is a merge, the allele in the merged population is sampled from the two merging populations x and y with respective probabilities p and $q = 1 - p$, where $p = \frac{N_{(i-1)x}(t_i-0)}{N_{(i-1)x}(t_i-0)+N_{(i-1)y}(t_i-0)}$. This results in the following formula for the joint distributions:

$$R_{ab}(t_i) = \begin{cases} R_{a'b'}(t_i - 0) & x \neq a',\ x \neq b' \\ pR_{a'b'}(t_i - 0) + qR_{yb'}(t_i - 0) & a' = x,\ b' \neq y \\ pR_{a'b'}(t_i - 0) + qR_{a'y}(t_i - 0) & b' = x,\ a' \neq y \\ p^2R_{xx}(t_i - 0) + 2pqR^+_{xy}(t_i - 0) + q^2R_{yy}(t_i - 0) & a' = x,\ b' = y \end{cases} \tag{6.5}$$

where $2R^+_{ab}(t) = R_{ab}(t) + R_{ba}(t)$.

A single migration event from one population x to another y can be seen as a merge of the whole destination population y with a part of the population x. Only the distributions of the destination population are affected. Assuming that the event occurred at time t_i is described by the migration matrix $M(t_i)$, the size of the part of population x contributing to the event is given by $m_{xy}(t_i)N_{(i-1)x}(t_i - 0)$.

A migration event in the network describes all possible migrations between two populations from the network. Therefore, a single population may be affected by many different migrations taking place at the same time. It leads to very complex relationships between joint distributions characterizing populations involved in the migration waves. The simplest way to model such a migration event relies on the following two-step scenario:

- Split each population from the network $\kappa_i - 1$ times in order to differentiate all migrating subpopulations. The population size ratio parameters used in these splits are given by the migration matrix $M(t_i)$.

- Merge all migrating subpopulations isolated in the previous step with proper destination populations. It is not necessary to apply any particular order to these merges.

As we see, the described method requires κ_i^2 populations (and κ_i^4 joint distributions) to be stored at the same time. We can optimize this method by reordering the scenario of splits and merges. For example, we can firstly isolate the joint migrating subpopulations from each original population, and then, apply the splits and merges scenario κ_i times, once for each isolated joint subpopulation. Each merge operations should immediately follow the split. It leads to the necessity of managing at most $2\kappa_i + 1$ populations at the same time.

However, if we assume that the migration rates are small, we can discard the relationships between small migrating subpopulations. It allows us, with the cost of introducing a small error to the results, to use much simpler computations to obtain the $R_{ab}(t_i)$ distributions after the migration event. We do so by treating each single migration event as a merge event of a part of original population with the whole destination population. In that case, as we see from (6.5), the joint distribution $R_{ab}(t_i)$ is modified by migrations from any population x, $x \in 0, 1, \ldots, \kappa_{i-1}$, provided that $m_{xb}(t_i) > 0$. Hence, based on (6.5), we calculate κ_i distributions $R_{ab}(t_i)$. Each of these distributions corresponds to the case that the only migration at time t_i to the population b took place from the population $x \in 0, \ldots, \kappa_{i-1}$ and we denote these distributions by $R_{ab}^{(x)}(t_i)$. Each migration from x to b changes the value of the joint distribution by $C_{xab}(t_i) = R_{ab}^{(x)}(t_i) - R_{ab}(t_i - 0)$. Then, we obtain the following formula for the joint distributions under the migration event:

$$R_{ab}(t_i) = R_{ab}(t_i - 0) + \sum_{0 \leq x < \kappa_i} C_{xab}(t_i). \tag{6.6}$$

We advice to use the optimized version of the first method unless the migration rates are very small or the allelic space is very large.

The user may specify if the migration event changes the population sizes. In that case the modified values of the population size satisfy the following formula:

$$N_{ix}(t_i) = \Big(1 - \sum_{k=0}^{\kappa_{i-1}} m_{xk}(t_i)\Big) N_{(i-1)x}(t_i - 0) + \sum_{k=0}^{\kappa_{i-1}} m_{kx}(t_i) N_{(i-1)k}(t_i - 0) \tag{6.7}$$

6.2 Expression for evolution of a joint distribution of a pair of individuals randomly sampled under any Markov mutation model

Current section follows derivations in Bobrowski et al. (19). We assume that each chromosome evolves under genetic drift and mutation between two consecutive network events. We assume that the mutation model is a time-homogenous Markov mutation. Accordingly, the allelic state $X_a(t) \in \mathbb{A}$ of the chromosome sampled from population a at time t evolves as a continuous-time non-negative Markov chain with transition intensity matrix $Q_a = \{q_{jk}^{(a)}\}, 1 \leq j, k \leq N_{\mathbb{A}}$, where $q_{jk} \geq 0, j \neq k$ and $\forall_j \sum_k q_{jk} = 0$. Thus, $r_{ab}[j,k](t) = P[X_a(t) = j, X_b(t) = k]$, where $j, k \in \mathbb{A}$ and $0 \leq a, b < \kappa_i$ if $t \in [t_i, t_{i+1})$. We assume that the matrix Q_a stays unchanged between two demographic events but may vary between different populations or different time intervals (the state space of the chain remains the same). By $P_a(t) = \{p_{jk}^{(a)}(t)\}, 1 \leq j, k \leq N_{\mathbb{A}}$ we denote the probability transition matrix corresponding to the matrix Q_a. In the finite-dimensional case (if $\mathbb{A}$ is finite) we obtain $P_a(t) = \mathrm{e}^{Q_a t}$.

Let us assume that the MRCA of two randomly chosen individuals with allelic types a_j and a_k ($a_j, a_k \in \mathbb{A}$, $1 \leq j, k \leq N_{\mathbb{A}}$) existed at time T_{jk} in the past, for example, before the present time t. Given population size $N(t)$ we obtain that $P[T_{jk} > \tau] = \mathrm{e}^{-\int_0^\tau N(t-u)\mathrm{d}u}$. The MRCA can be of any allelic type a_i with index i (each one with probability $\pi_i(T_{jk}) = P[X(T_{jk}) = a_i]$) and its descendants at the present time t have types to a_j and a_k, respectively. Then, summing over all possible values of i and following (19), we obtain:

$$r[a_j, a_k](t) = \int_{-\infty}^{t} \sum_{1 \leq i \leq N_{\mathbb{A}}} \pi_i(\tau) p_{ij}(t-\tau) p_{ik}(t-\tau) \frac{1}{N(\tau)} \mathrm{e}^{-\int_\tau^t \frac{\mathrm{d}u}{N(u)}} \mathrm{d}\tau \tag{6.8}$$

Expression (6.8) may be transformed into matrix notation and we can separate the evolution of the population in the time interval before $t = 0$ and interpret it as the initial conditions (19). This leads to the following equation:

$$R(t) = P^T(t) R(0) P(t) \mathrm{e}^{-\int_0^t \frac{\mathrm{d}u}{N(u)}} + \int_0^t P^T(t-\tau) \Pi(\tau) P(t-\tau) \frac{1}{N(\tau)} \mathrm{e}^{-\int_\tau^t \frac{\mathrm{d}u}{N(u)}} \mathrm{d}\tau, \tag{6.9}$$

where P^T is the transpose of the matrix P and $\Pi(t)$ is a diagonal matrix with $(\Pi(t)_{ii}) = \pi_i(t)$.

$R(t)$ given by expression (6.9) is a mild solution (150) of the following matrix differential equation known as the Lyapunov equation (54):

$$\frac{\mathrm{d}R_{ab}(t)}{\mathrm{d}t} = Q_a^T R_{ab}(t) + R_{ab}(t)Q_b + \frac{\delta_{ab}}{N_a(t)}(\Pi(t) - R_{ab}(t)), \tag{6.10}$$

where $t \in [t_i, t_{i+1}), 0 \leq a, b < \kappa_i$ and δ_{ab} is the Kronecker delta. This equation has a unique solution of the form of:

$$R_{ab}(t) = \begin{cases} P_a^T(t-t_i)R_{ab}(t_i)P_b(t-t_i)\mathrm{e}^{-\int_{t_i}^{t}\frac{\mathrm{d}u}{N_a(u)}} + S_a(t_i,t) & a = b \\ P_a^T(t-t_i)R_{ab}(t_i)P_b(t-t_i) & a \neq b, \end{cases} \tag{6.11}$$

where $S_a(t_i,t) = \int_{t_i}^{t} P_a^T(t-\tau)\Pi(\tau)P_a(t-\tau)\frac{1}{N_a(\tau)}\mathrm{e}^{-\int_{\tau}^{t}\frac{\mathrm{d}u}{N_a(u)}}\mathrm{d}\tau$

We will use (6.10) rather than (6.11) to calculate the evaluation of the joint distribution in the time interval between two adjacent events in the network.

6.3 Model refinements

6.3.1 Sample of size greater than 2

Formula (6.8) expresses the joint distribution of the pair of individuals randomly sampled from the population. Unfortunately, as it follows from the de Finetti's theorem (35, 160), we cannot directly apply the values of this joint distribution to obtain results for a larger sample. Moreover, the joint distribution of all individuals from a sample of size n is a n-dimensional array with the total number of entries equal to $N_{\mathbb{A}}^n$. Nevertheless, we can obtain this joint distribution for small values of n and $N_{\mathbb{A}}$.

We assume that each individual from the population is represented by a haplotype sequence with two possible nucleotides at each position. We denote the joint probability of n individuals (ith individual being of type $a_i, 1 \leq i \leq n, 1 \leq a_i \leq N_{\mathbb{A}}$) randomly sampled from population at time $t \in [t_i, t_{i+1})$ by $r^{(n)}_{a_1 \ldots a_n}(t)$. Let us assume that the MRCA of a pair of individuals from the given sample existed at time τ and that a_j and a_k ($1 \leq j, k \leq n, j < k$) are the direct descendants of this MRCA in the current generation. We describe the sample at time $\tau - 0$ just before this coalescent event by $b_1 \ldots b_{n-1}$, where the ith individual becomes at time t either the individual of type a_i if $i < k$ or of type a_{i+1} otherwise. The probability that individual of type x at time τ

evolves into individual of type y at time t is given by the transition probability value $P_{xy}(t-\tau)$. Thus, summing over all possible values of j, k and $b_i, 1 \leq i < n$, we obtain:

$$\begin{aligned} r^{(n)}_{a_1 \ldots a_n}(t) = \int_{-\infty}^{t} \sum_{b_1 \ldots b_{n-1}} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} \frac{r^{(n-1)}_{b_1 \ldots b_{n-1}}(\tau)}{\binom{n}{2}} \prod_{i=1}^{k-1} P_{b_i a_i}(t-\tau) \cdot \\ \cdot P_{b_j a_k}(t-\tau) \prod_{i=k}^{n-1} P_{b_i a_{i+1}}(t-\tau) \binom{n}{2} \frac{\mathrm{e}^{-\binom{n}{2} \int_{\tau}^{t} \frac{\mathrm{d}u}{N(u)}}}{N(\tau)} \mathrm{d}\tau \end{aligned} \tag{6.12}$$

The last step is to calculate the values of $P_{xy}(t)$, $1 \leq x, y \leq N_{\mathbb{A}}$. Assume that haplotype sequences contain only one nucleotide ($s = 1$). Then, the mutation intensity matrix is of the following form:

$$Q = \begin{vmatrix} -\mu & \mu \\ \mu & -\mu \end{vmatrix}, \tag{6.13}$$

where μ is the mutation rate.

For any matrix X, we define the matrix e^X by (6.14) (118).

$$\mathrm{e}^X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k \tag{6.14}$$

Therefore, in order to obtain the $P_{xy}(t) = \mathrm{e}^{Q_{xy}t}$ values, we firstly expand the P matrix into the power series according to (6.14):

$$P(t) = \begin{vmatrix} 1-\alpha & \alpha \\ \alpha & 1-\alpha \end{vmatrix} = \sum_{k=0}^{\infty} \frac{(Qt)^k}{k!} = \begin{vmatrix} 1 & 0 \\ 0 & 1 \end{vmatrix} + \begin{vmatrix} -\mu t & \mu t \\ \mu t & -\mu t \end{vmatrix} + \frac{1}{2!} \begin{vmatrix} 2\mu^2 t^2 & -2\mu^2 t^2 \\ -2\mu^2 t^2 & 2\mu^2 t^2 \end{vmatrix} + \ldots, \tag{6.15}$$

where α is the probability that the nucleotide changes the allele type in the time interval t. Expression (6.15) leads to the following formula for α:

$$1 - \alpha = \sum_{k=0}^{\infty} \frac{(-2)^{k-1}(\mu t)^k}{k!} = 0.5 + 0.5 \left(\sum_{k=0}^{\infty} \frac{(-2\mu t)^k}{k!} \right). \tag{6.16}$$

The sum from the right side of Formula (6.16) is the Maclaurin series of $\mathrm{e}^{-2\mu t}$. We generalize the case for any number of nucleotides s and using formulas (6.15) and (6.16) we obtain the final expression for $P_{xy}(t)$:

$$P_{xy}(t) = \left(0.5 + 0.5\mathrm{e}^{-2\mu t}\right)^{s-d(x,y)} \left(0.5 - 0.5\mathrm{e}^{-2\mu t}\right)^{d(x,y)}, \tag{6.17}$$

where $d(x, y)$ stands for the number of positions with different nucleotides in two individuals x and y.

Formula (6.17) has been verified for longer haplotype sequences in both ways, analytically (up to $s = 4$) and numerically by applying the main model for $s < 10$.

6.3.2 Model complexity reduction for some mutation models

The main factor that determines the complexity and feasibility of computations is the assumed model of mutation and, more precisely, the number $N_{\mathbb{A}}$ of all possible allelic types. Our approach works perfectly if we model a simple SNP mutation with not very large number of nucleotides or if we model microsatellites which rarely exceed one hundred tandem repeats in length. The problem appears when we try to approximate an infinite-site mutation model by applying our method for long haplotype (i.e., long SNP sequences). In this case $N_{\mathbb{A}} = z^s$, where s is the number of bases and z is the number of possible nucleotide variants at each base (usually $z = 2$). However, we can reduce the complexity of the problem under the assumption that the distributions of the mutation process are invariant under permutations of bases (exchangeable). We obtain the reduction in this case as follows.

Let us assume that $z = 2$ and the nucleotide at each base is either wild-type or mutated. We transform the allelic space $\mathbb{A}$, containing 2^s possible allelic types, into a much smaller $\mathbb{A}'$ set. For a given value of c, $1 \leq c \leq \lfloor s/2 \rfloor$, $\mathbb{A}'$ contains all possible pairs (a, b), where each pair (a, b), $0 \leq a \leq s$, $0 \leq b \leq a$ groups all allelic types from $\mathbb{A}$ with b mutated nucleotides at the first c bases and $a - b$ mutations at the bases from the $(c+1)$th to the sth. The total number of the (a, b) pairs depends on c and varies from $2s - 2$ for $c = 0$ to $s^2/4 + s + 1$ for $c = \lfloor s/2 \rfloor$. In order to retrieve the joint distribution for original $\mathbb{A}$ space, we need to review values of c in the range from 0 to $\lfloor s/2 \rfloor$. Table 6.1 presents an example of a $\mathbb{A}'$ set and the Q matrix for the case with $s = 4$, $c = 2$, with α being the mutation intensity at each base. We are able to obtain each value of the joint distribution of a pair of individuals from $\mathbb{A}$ by using the proper entry of the reduced joint distribution matrices divided by the total number of pairs grouped by this entry. For example, to calculate the joint probability rate of two individuals of the form of $XxxX$ and $XXXx$ (where X stands for a mutated nucleotide) we can use the matrix obtained for a case $s = 4, c = 2$ and the joint probability rate of entries $(3, 1)$ and $(2, 2)$ indicating the joint probability rate of two individuals with two and three

Table 6.1: Reduction of $\mathbb{A}$. The table presents the reduced mutation intensity (probability rate) matrix Q for the case of $s = 4$ and $c = 2$. The α value stands for a value of the mutation intensity at each base.

	$(0,0)$	$(1,0)$	$(1,1)$	$(2,0)$	$(2,1)$	$(2,2)$	$(3,1)$	$(3,2)$	$(4,2)$
$(0,0)$	-4α	2α	2α	0	0	0	0	0	0
$(1,0)$	α	-4α	0	α	2α	0	0	0	0
$(1,1)$	α	0	-4α	0	2α	α	0	0	0
$(2,0)$	0	2α	0	-4α	0	0	2α	0	0
$(2,1)$	0	α	α	0	-4α	0	α	α	0
$(2,2)$	0	0	2α	0	0	-4α	0	2α	0
$(3,1)$	0	0	0	α	2α	0	-4α	0	α
$(3,2)$	0	0	0	0	2α	α	0	-4α	α
$(4,2)$	0	0	0	0	0	0	2α	2α	-4α

mutations, respectively, including exactly one common mutation. At the end we need to divide the value of the joint probability rate by 2 because there are exactly two ways to represent a $(3,1)$ entry (as $XxXX$ or as $xXXX$) and only one way to represent a $(2,2)$ entry (as $XXxx$).

The reduction method described in this section allows us to obtain the results in a reasonable amount of time even for a model with $s \approx 100$. We may also use this reduction for $z > 2$ but it requires a more complex enumeration of the states of the $\mathbb{A}'$ set. These divisions would be multi-dimensional, compared to the two-dimensional case explained above, because of the necessity of taking into account the exact types of nucleotides being changed in the mutation process.

6.4 Model implementation

The program attached to the thesis calculates the joint distribution in a demographic network described by an input script file. Along with an executable version of the program we also attach a complete source code including a few additional auxiliary modules that improve the basic functionality of our program. We describe the program in more details further in this section. Finally, we add an implementation of (6.12) that allows to calculate the joint distribution of individuals based on a sample of size greater than two. We use Newton-Cotes method (5) with Boole's rule (176) for numerical

integration given by formula (6.12). Despite of the complex recurrence dependency and the multidimensional form of the joint distribution, we are able to obtain results for small $n > 2$ and simple mutation models.

6.4.1 Program structure

The program simulates demography of a set of populations and calculates the joint distribution of individuals from these populations according to a given input file (for more details about input script file see Section 6.4.4). Thus, in the default version the main part of the program analyzes the input file and executes a single experiment. It may be easily changed in order to run more complex calculations, such as the reduction of the model described in Section 6.3.2. Besides that, the program consists of six modules:

1. Module Model contains description of the network and handles events occurring in the network along with evolution of populations between the events.
2. Module Population contains the information about size and growth scenario of a single population. In the default version we can choose either constant or exponential growth of population size but we can also easily add any other model of growth.
3. Module JointDistr stores a single joint distribution of individuals between any two population.
4. Module Matrix implements our own representation of a matrix accomplishing multiplication of sparse matrices in 2nd-order polynomial time complexity.
5. Auxiliary module Methods allows to add procedures that compute model parameters based on the obtained joint distributions.
6. Auxiliary module DataFile helps to create an input script file.

6.4.2 Algorithms

We store a set of populations as a vector and their joint distribution as a list of matrices. Each discrete event in the network is implemented according to Formulas (6.4-6.6). The chosen data representation allows fast and simple managing of population's indices.

Mutation and drift effect in the time interval between two events is computed according to the Lyapunov equation (6.10). In order to numerically solve the ODE we use the RK4 algorithm (Runge-Kutta 4th order method) (53, 55), with adaptive control of the step size using the Cash-Karp method (22, 158).

The intensity matrix Q for large number of allele types $N_{\mathbb{A}}$ is usually a sparse matrix. Moreover, the most time-consuming operation of the implemented Runge-Kutta algorithm for our ODEs is the RQ multiplication. Thus, a proper matrix representation, taking into account a fact that usually the matrix Q is a sparse matrix, significantly improves the time complexity. We accomplish this by storing, for a matrix with the percentage of non-zero entries not exceeding a limit value, all non-zero values in a set of lists. Each list corresponds to a single row or column of the matrix. This structure allow to achieve 2nd-order polynomial time-complexity of multiplication operation if at least one of the multiplying matrix is a sparse matrix.

6.4.3 Time and memory complexity

Both time T and memory M complexities depend on the number of populations n and the size of the allelic types space $N_{\mathbb{A}}$. The time complexity also depends on the form of the mutation model. We assume that the intensity matrix Q is sparse with the average of $c \ll N_{\mathbb{A}}$ nonzero values per row (or per column). In each RK4 step we need to execute exactly 12 matrix multiplication, each one with complexity $cN_{\mathbb{A}}^2$ and about 60 other operations running in $cN_{\mathbb{A}}$ time but requiring the initialization of the matrix. Thus, $T = \kappa^2 kr(60 + 8c)N_{\mathbb{A}}^2$, where k is a number of splits or merges and r is the average number of steps in the RK4 algorithm for a single time interval (usually $r < 100$, especially when we use an adaptive step control algorithm). The method is feasible even for $N_{\mathbb{A}} \approx 1000$ and $\kappa > 10$.

In the algorithm we need to store n intensity matrices and n^2 joint distributions, therefore $M = 8(n^2 + n)N_{\mathbb{A}}^2$ [Byte]. As we see, the memory limit should not be a problem even for the largest feasible cases.

6.4.4 Sample input script

The program requires a single input file that describes the network. Since we assume that we start with a single population at time $t = 0$, the file should contain the initial

joint distribution of individuals from this ancestral population along with the population size and growth scenario and the description of the mutation model given by the intensity matrix Q. The model of mutation is constant but intensities may vary among different time intervals or different populations. To simulate evolution in time, we define the entry in the input file that allows us to move from the current to the given generation. Thus, all events in the input file should be ordered chronologically. Algorithm 6.1 is a sample input script for a demographic network with a single split event and a simple SNPs mutation model.

Algorithm 6.1 Sample input script file.

```
mut 0 2 2          //mutation model and initial Q matrix, N_A = 2
-0.00000125 0.00000125
0.00000125 -0.00000125
r0                 //initial joint distribution
0.36 0.24
0.24 0.16
ps 0 4200          //set the population size of population 0
s 0 0 0.5          //split population 0
g 4300             //move in time
pg 0 e 0.00106     //population 0 grows exponentially
pg 1 e -0.00585    //population 1 shrinks exponentially
g 5040             //move in time
pg 0 e 0.00880     //change the growth speed of population 0
g 6000             //move in time
pf out.txt         //store the results
```

6.5 Sample applications

6.5.1 Equilibrium estimates

We model a population with constant size $N = 1000$. We consider two individuals from the population and check allelic types at single specific homologous SNP in these individuals (being one of the two nucleotides A or a that can occur at that SNP). We also assume that each nucleotide can mutate at the rate 0.002 per generation. The

population is in a mutation-drift equilibrium and using our program we obtain that in equilibrium the values of the joint distributions are equal to: $r[A,A](t) = r[a,a](t) = \frac{5}{18}$ and $r[A,a](t) = r[a,A](t) = 0.5 - r[A,A](t) = \frac{2}{9}$. These values may also be obtained using asymptotic analysis of the Lyapunov equation (6.10).

Next, we assume that at time $t = 0$ the population splits into two populations. The first population, with index 0, is the ancestral population. The second population, with index 1, starts with 1000 individuals and grows exponentially with a parameter 0.001. Mutation rate in both population is unchanged. We want to analyze how the demography affects the association between individuals from the same or different populations. For this purpose we compute the value of the normalized Lewontin linkage disequilibrium D' between populations i and j (119): $D'(t) = D(t)/D_{max}(t)$, where $D(t) = r_{ij}[A,A](t) - p_1 q_1$ is a non-normalized linkage disequilibrium and $D_{max}(t) = \min(p_1 q_1, p_2 q_2)$ for $D(t) < 0$ or $D_{max}(t) = \min(p_1 q_2, p_2 q_1)$ otherwise. Lewontin's index is usually applied to quantify dependence (linkage) between alleles of different loci of the same chromosome. Here it is used to quantify dependence between the alleles at the same locus of different chromosomes. By p and q we denote the constraint distributions of the joint distributions as follows: $p_1 = r_{ij}[A,A](t) + r_{ij}[A,a](t)$, $p_2 = r_{ij}[a,A](t) + r_{ij}[a,a](t)$, $q_1 = r_{ij}[A,A](t) + r_{ij}[a,A](t)$ and $q_2 = r_{ij}[A,a](t) + r_{ij}[a,a](t)$. The results presented in Figure 6.1 are intuitively clear. The joint distribution in population 0 stays constant while in population 1 it slowly evolves leading to a decrease of the value of D' – the force of drift decreases in a growing population. The common association of homologous loci from two different populations after a split event rapidly decreases.

6.5.2 Predictions and estimates of a common species and populations history

Availability of the genetic data from different species and populations allows for various, often very sophisticated, intra- and inter-population analysis. Particularly, it is possible to estimate a past demography of these populations, including interactions between populations. The most common method to do so involves using the model-based analysis (170). In this approach we may consider one or more past demographic scenarios and either test or adjust them based on a given data. As an example of

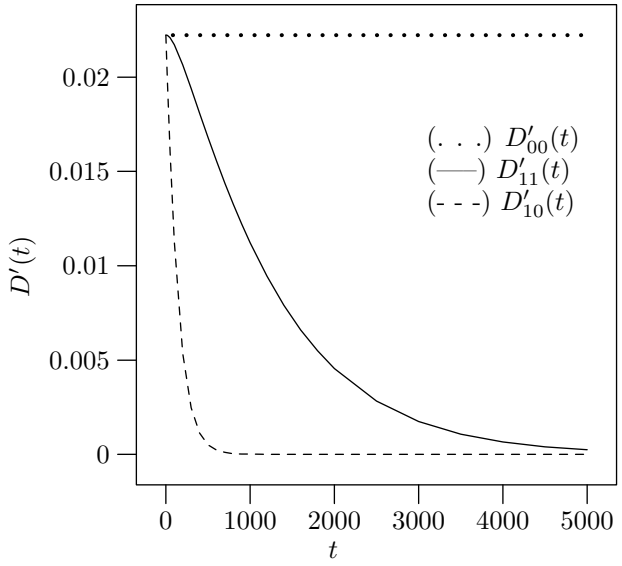

Figure 6.1: The Lewontin's index in a constant size population (D'_{00}), an exponentially growing population (D'_{11}) and between these two populations (D'_{10}). Both populations evolved from a common ancestral population, the split event occurred at time $t = 0$. The ancestral population was in a mutation-drift equilibrium.

such a model we can mention works of the Barbujani's group who studied the genetical relationship between Etruscans and Tuscans (14) or between Neandertals (N), Cro-Magnoid (CM) and modern Europeans (M) (13). In both papers the authors calculated the values of several parameters (such as pairwise difference or haplotype diversity) for samples drawn from populations simulated under about dozen hypothetical demographic scenarios and compared them to the data obtained from the real individuals. Our demographic network model perfectly suits this kind of approaches. We apply our method to the scenarios used by Barbujani's group in the second paper in order to obtain the estimates of the pairwise difference between populations. The models used in our calculations are listed in Table 6.2. All of these models assume that N and CM lived 1700 and 960 generations ago, respectively. All models except L1.7 assume a single population with different growth rates. Model L1.1 is a constant size population. In L1.2, the population grows after origin of CM. Models L1.3 and L1.4 introduce to L1.2 a small growing rate before origin of N; rapid expansion from CM to M is assumed in L1.4. In L1.5 the population grows to the same large size as in L1.4, but with more balanced growing rates in each time interval. L1.7 models the population

Table 6.2: Pairwise difference calculations. The table presents the values of pairwise difference obtained by applying two different methods (simulation approach and our method based on the demographic network) to three populations: Cro-Magnoid (CM), Neanderthal (N) and modern European (M) under different demographic scenarios (explained in Belle et al. paper). Infinite-site model, approximated by haplotype sequences consist of 360 nucleotides, was assumed. Simulated approach required a fixed sample size to be specified for each population (N – 6, CM – 2 and M – 558). The table presents median values of the pairwise difference for simulation method and mean values for our method.

Method	Population	Demographic model								
		L1.1	*L1.2*	*L1.3*	*L1.4*	*L1.5*	*L1.7*	*H1.1*	*H1.2*	*H1.3*
Belle	N	1.9	1.9	0.9	1.5	12.9	3.9	15	7	4.7
	CM	1	1	1	1	11	4	12	10	7
	M	1.7	2.4	1.9	2.3	13.6	4.7	18.6	14.6	13
Our	N	2.2	2.2	1.1	1.7	17.1	1.9	20	7.8	5.2
	CM	2.2	2.2	1.4	1.8	17.3	2.6	20	11.3	8.5
	M	2.2	2.8	2.1	2.6	17.9	3.4	25.5	17.9	15.6

from L1.5 with an assumption that there existed a separated shrinking N population. Models starting with H have ten times larger mutation rate assumed (0.5 per million years per nucleotide instead of 0.05 as in models staring with L). Demography of H1.1 is the same as in L1.2. Demographies of H1.2 and H1.3 are slightly different versions of L1.3. For more details see (13). Long haplotype sequences of 360 nucleotides with two possible variants at each position are assumed. Given the joint distribution $R_{xx}(t)$ we can calculate the mean pairwise difference $\varphi_x(t)$ in population x according to the following formula:

$$\varphi_x(t) = \sum_{i,j \in \mathbb{A}} d(i,j) r_{xx}[i,j](t), \tag{6.18}$$

where $d(i,j)$ is the number of positions on which the sequences of allelic types i and j differ. Table 6.2 compares our results to those obtained by Belle and co-workers in (13). Discrepancies between simulation and our methods, which are the largest for the CM population, can be explained by a very small CM sample size (only 2 individuals), and the fact that median (not mean) value was listed by Belle et al. The sample sizes of N and M populations used in the simulation approach were equal to 6 and 558, respectively.

6. DEMOGRAPHIC NETWORK MODEL

We use the demographic network model to study a common history of several Eastern European populations based on the Y chromosome data. We are interested in the history of relations between the Slavs occupying territory of the modern Poland and the Central Balts – ancestors of the modern Lithuanians and Latvians. Both the Proto-Slavs and the Balts are likely to have originated from nomad tribes that had left the Indo-European homeland and gave birth to the most of the modern European cultures. It is argued that both these groups came to Europe in the 2nd millenium BC (20, 62, 70), but the exact chronology is unclear. Both populations, Polish and Lithuanian, are considerably genetically distinctive (156) from other European nations. However, analysis of the Y-chromosome haplogroups of the modern Eastern European citizens show that other ancient populations should also be considered (69). The Baltic-Slavic branch of the Indo-European genealogy belongs to the R1a haplogroup. However, about 45% of the population of the modern Lithuania and Latvia belong to the N1c1 haplogroup, which is a haplogroup of the North-European ancestors (the Finns) that entered Europe during the Corded Ware period at the beginning of the 3rd millenium BC (10). The Balts, after the split from other Indo-European groups, settled in the areas along the south-eastern coast of the Baltic Sea and assimilated with the Finn tribes living there. Our studies confirms that the influence of the N1c1 haplogroup cannot be discarded from the Slavic-Balts analysis; the results obtained from the model that considers only the Balts and the Slavs indicate much closer relationship between these two groups than it follows from the genetic data. The exact times of splits of the Balts and the Slavs, or the Finns from other Indo-European tribes, are unknown. Therefore, estimating them becomes one of the aims of our studies. Another interesting aspect of the Balt-Slav relationship concerns migration waves that took place between these two groups in the 6th, when the Slavs appeared on the territory of Poland for the first time, and in the 14th century, when the Commonwealth of Poland and Lithuania was formed. Figure 6.2 illustrates the modeled demographic scenario. We estimate the population sizes of all groups based on Kremer's population growth rates (113) as described in Section 5.4.2. Since the effective to census population size ratio in humans is estimated to be equal between 0.3 (101) and 0.5 (140), and since considerating the Y-STR haplotype chromosome introduces a factor 1/4 to this ratio (154), we assume that the effective population size is ten times smaller than the census data size.

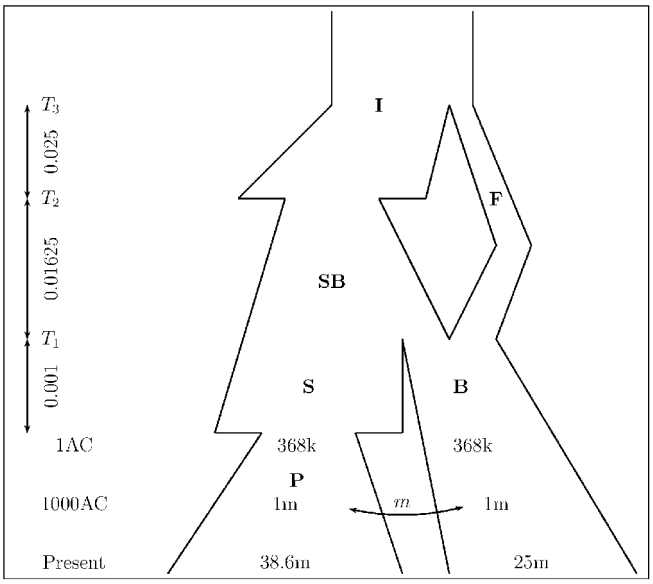

Figure 6.2: Balt-Slav-Finn demographic model. The model includes the following populations: Indo-Europeans (I), Finns (F), Slavs (S), Balts (B) and Poland (P). Population sizes before the year 1AC are estimated using Kremer's rates (the values listed on the left – growth rates per generation with assumption that a single generation lasts 25 years). Two bottleneck events indicate: (i) S and B tribes leaving the Indo-European homeland and (ii) isolation of P from S. The strength of the bottlenecks only slightly changes the value of R_{ST} provided that the population after the bottleneck has a size equal at most of 1/3 of the original size. Thus, we assume the following values of the bottleneck size ratios: 1/5 in the case of SB and 1/7 for P. Four parameters are varied: (i) T_1 – time of split of B and S, (ii) T_2 – time of BS leaving the Indo-European homeland (iii) T_3 – time of split of I and F, and (iv) m – migration rate between B and P. The exact time of the migration does not influence the R_{ST} value.

We use the R_{ST} Slatkin's distance (168) to quantify distance between two populations. We calculate the R_{ST} distance between the Slavs and the Balts based on the data from 919 unrelated male Polish individuals sampled from six geographical regions of Poland and 297 Balts descendants (152 from Vilnius and 145 from Riga). Genetic data at nine microsatellite loci is considered: DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a and DYS385b. The data can be obtained from (103) or (94). We use the Arlequin program (46) to obtain normalized (61) value of the Slatkin's distance. We obtain that the R_{ST} value for samples of Poles and Balts is equal to 0.03862.

Our aim is to adjust the parameters of the scenario presented in Figure 6.2 in order to obtain a realistic model explaining obtained value of the R_{ST} distance. We model differences in the number of tandem repeats between two loci rather than the exact numbers themself in the mutation model in the demographic network. Therefore, we substitute the Lyapunov equation (6.10) by the following specialized equation:

$$\frac{\mathrm{d}R_{ab}(s,t)}{\mathrm{d}t} = -(v_a + v_b)\Big(1 - \frac{s}{2} - \frac{1}{2s}\Big)R_{ab}(s,t) + \frac{\delta_{ab}}{N_a t}\Big(\Pi - R_{ab}(s,t)\Big), \tag{6.19}$$

where $R_{ab}(s,t) = \sum_{i=-\infty}^{\infty} s^i r_{ab}(i,t)$, is the probability generating function of an integer-valued random variableequal to the difference in allele size between individuals sampled from populations a and b ($0 \leq a, b < \kappa_i$) at time $t \in [t_i, t_{i+1})$, v_a and v_b are mutation intensities in both populations, Π is a vector of the same size as $R(s,t)$ with the value of 1 in the middle and all other entries equal to 0 and δ_{ab} is the Kronecker delta. Detailed description regarding obtaining expression (6.19) may be found in (107). Given the values of $R_{aa}(s,t)$, $R_{bb}(s,t)$ and $R_{ab}(s,t)$, one can calculate the average sum of squared difference distance (R_{ST} distance) between populations a and b by using the following formula:

$$R_{ST} = \frac{2V_{ab}(t) - V_{aa}(t) - V_{bb}(t)}{V_{ab}(t)}, \tag{6.20}$$

where $V_{xy}(t)$ is the variance of the allele size in joint populations x and y ($x = y$ is admissible) given by the following formula:

$$V_{xy}(t) = \sum_{i=-\infty}^{\infty} i^2 r_{xy}(i,t) \tag{6.21}$$

Figures 6.3-6.5 present estimates of the genetic distance between descendants of the Slavs and the Balts for different times of splits and migration rates. As we see, the required value of the the Slatkin's distance may be obtained by more than one set of parameter's values. Therefore, although the estimates deliver useful information about common history of modeled groups, they are not sufficient to determine the exact scenario of the past demography. The other important issue that one needs to be aware of is that these estimates are strongly dependent on the effective population sizes, which cannot be precisely estimated. However, more complex analysis of the allele size distributions and replacing the genetic drift model by the genetic draft model (59),

which is unaffected by the effective population size, should suffice to overcome these problems. We leave these investigations for further studies.

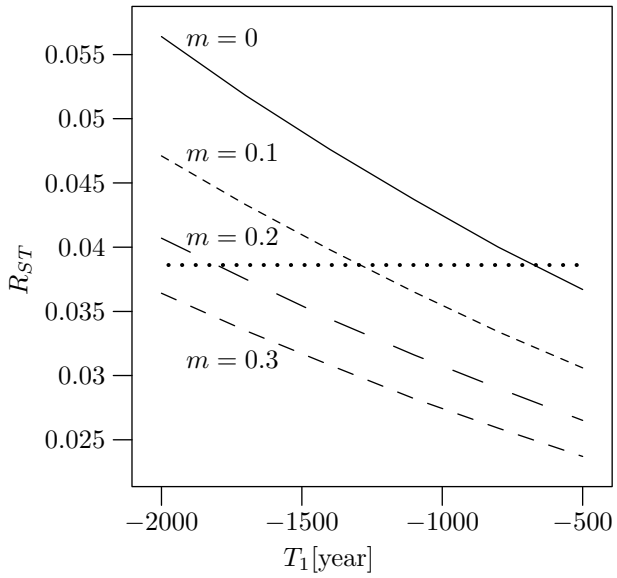

Figure 6.3: R_{ST} distance between the Slavs and the Balts as a function of the Balt-Slav split time T_1. Populations evolve according to the model presented in Figure 6.2 with $T_2 = -3500$ and $T_3 = -15000$. Results for four different values of the migration rate m are depicted. Dotted line indicates the real data estimate of R_{ST}.

6.5.3 Ascertainment bias model for microsatellite loci

Microsatellite loci, due to their very high polymorphism, are commonly used as genetic markers. Because of the predominant model of mutation (extensions and contractions of the repeat sequence), a high polymorphism at a particular locus usually manifests itself by presence of long repeat sequences at this locus. Thus, when we compare variability between two species, we often analyze these microsatellite loci. When we choose highly polymorphic loci based on the studies of species 1 and then compare the average number of repeats at loci homologous ih species 1 and 2, we usually obtain that sequences from species 1 are longer (30). This phenomenon can be explained by a tighter correlation of the allele sizes between homologous loci in one species than in two different species. We call this bias the ascertainment bias. Thus, when we obtain differences of lengths of homologous loci from two populations, important question arises: Is the ascertainment

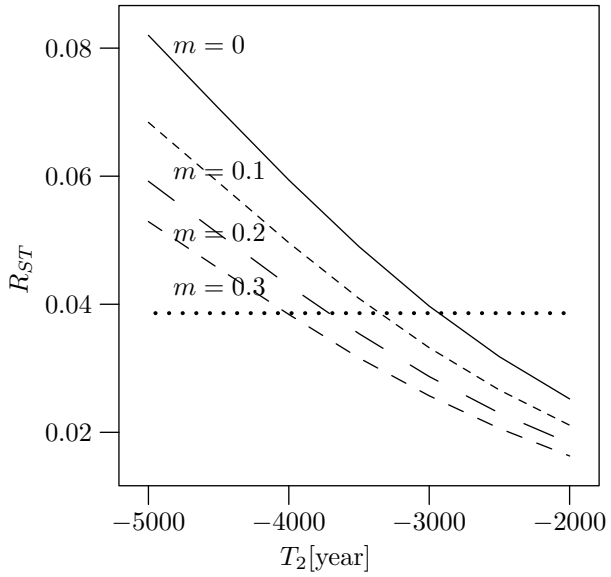

Figure 6.4: R_{ST} distance between the Slavs and the Balts as a function of the time T_2 when the Balts and the Slavs left the Indo-European homeland. Populations evolve according to the model presented in Figure 6.2 with $T_1 = -1500$ and $T_3 = -15000$. Results for four different values of the migration rate m are depicted. Dotted line indicates the real data estimates.

bias the sole explanation for this difference or not ? Recent studies show that these differences between many species (i.e., between human and chimpanzee) may result also from different demographies and mutation processes (28, 177). Therefore, it is important to develop a method that can distinguish the effect of the ascertainment bias from evolutionary factors. In the most basic method we obtain the results for two scenarios. In each of them we consider polymorphic loci discovered in the other population and at the end we compare these results. This approach is known as the reciprocal study (41).

There are several methods that allow to estimate the value of ascertainment bias (104). We model the ascertainment bias in the following way. Let us assume that ancestral population (the common ancestor of both species that we analyze) evolves from time $t = 0$ with the initial number of repeats equal to 1. At time $t = T_0$ the population splits into two species. Then, both species evolve until the current time $t = T$. We denote the effective population sizes as N_0 for ancestral species and N_1 and

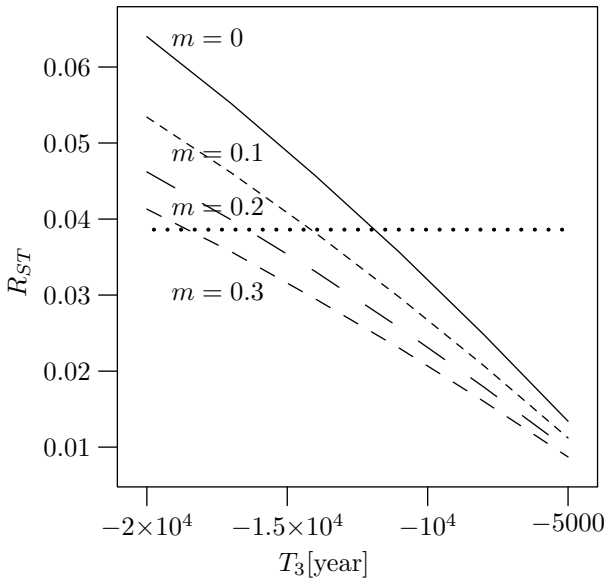

Figure 6.5: R_{ST} distance between the Slavs and the Balts as a function of the Finn-Indo-European split time T_3. Populations evolve according to the model presented in Figure 6.2 with $T_1 = -1500$ and $T_2 = -3500$. Results for four different values of the migration rate m are depicted. Dotted line indicates the real data estimates.

N_2 for the first and the second current species, respectively. At time $t = T$ we choose a locus with the allele length Y_0 greater than x from the first species and compare it with allele lengths at the same locus, obtained independently from both species. We denote these lengths by random variables Y_i, where $i \in 1, 2$ is the species index. Finally, we get the following formula for the expected value of the ascertainment bias B:

$$B = \mathrm{E}[Y_1|Y_0 > x] - \mathrm{E}[Y_2|Y_0 > x] = \frac{\sum_{i \geq x}(\mathrm{E}[Y_1, Y_0 = i] - \mathrm{E}[Y_2, Y_0 = i])}{\sum_{i \geq x} P(Y_0 = i)}. \tag{6.22}$$

The value of B given by (6.22) is computed from the joint distributions $R_{00}(T)$ and $R_{10}(T)$, where population 0 is the one from which we choose polymorphical loci:

$$B = \frac{\sum_{i,j \in \mathbb{A}, j \geq x} i r_{00}[i,j](T) - \sum_{i,j \in \mathbb{A}, j \geq x} i r_{10}[i,j](T)}{\sum_{i,j \in \mathbb{A}, j \geq x} r_{00}[i,j](T)} \tag{6.23}$$

Our demography model can serve as the model of the human-chimpanzee genealogy. We consider recent $T = 5 \cdot 10^5$ generations (average generation lifetime is assumed equal to 20 years). Thus, the split occurred at time $T_0 = 3 \cdot 10^5$ generations (85). We assume

the effective population sizes as $N_0 = 5 \cdot 10^4$ for ancestral common population (85), $N_1 = 1 \cdot 10^4$ for human (174) and $N_2 = 3 \cdot 10^4$ for chimpanzee (86). We assume that during each realization of a single experiment the mutation model is a SMM model with constant rate $v_i, i \in 0, 1, 2$ for the ith species, which is not entirely accurate – in fact longer loci may have a higher mutation rate (182), and asymmetry parameter $b = 0.55$. This latter has been introduced to reduce the number of loci with the number of tandem repeats decreasing to 0. We also set $x = 12$.

Figures 6.6 and 6.7 presents how the estimation of B depends on mutation rate and population sizes.

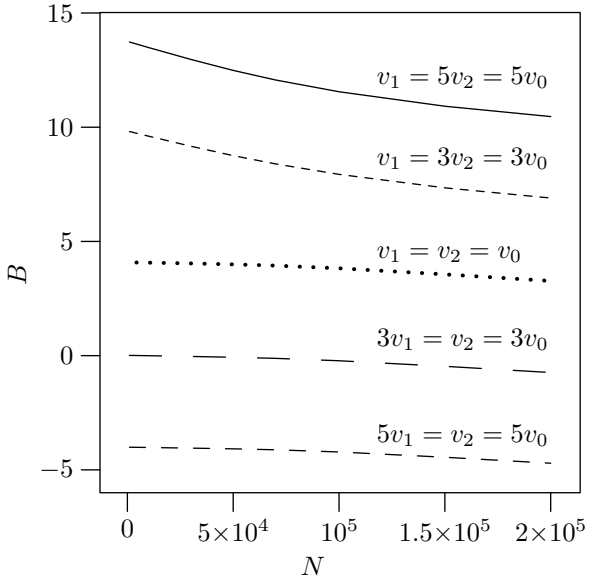

Figure 6.6: Estimation of the ascertainment bias B as a function of the mutation rate. We assume that $T_0 = 3 \cdot 10^5$, $T = 5 \cdot 10^5$ generations, $N_0 = N_1 = N_2 = N$ and $v_0 = 10^{-4}$ mutations per generation.

Figure 6.9 shows reciprocal studies of the human-chimpanzee relationship. As we can observe a marked difference in the differences between the average lengths of homologous loci sampled from human and chimpanzee under two scenarios in which we choose loci either based on human data or chimpanzee data (28). This fact suggests that the human's microsatellite mutation rate is higher.

In order to explain the ascertainment bias phenomenon we should study the evolution of microsatellites. We notice that ascertainment bias should not exist if a mi-

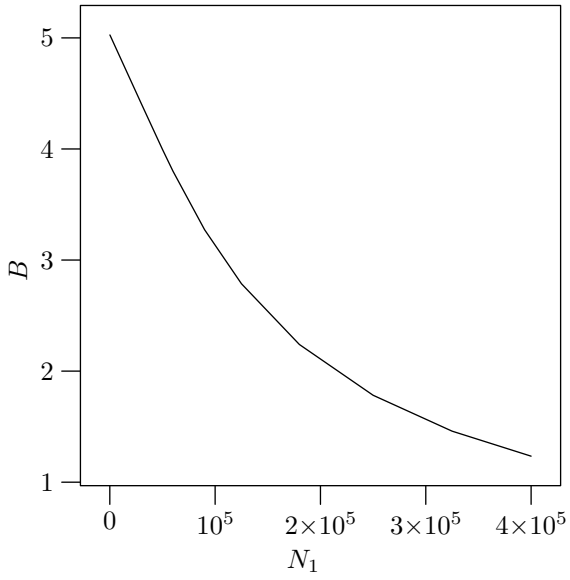

Figure 6.7: Ascertainment bias B as a function of the population size. We assume that $T_0 = 3 \cdot 10^5$, $T = 5 \cdot 10^5$ generations, $N_0 = 5 \cdot 10^4$, $N_2 = 3 \cdot 10^4$ and $v_0 = v_1 = v_2 = 10^{-4}$ mutations per generation.

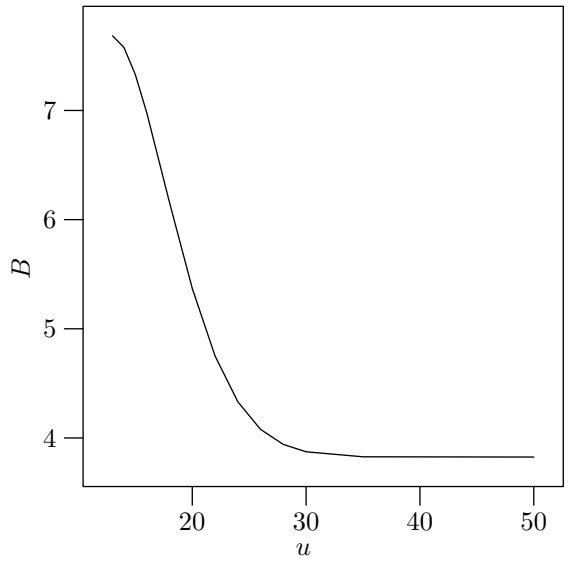

Figure 6.8: Estimation of the ascertainment bias B with the upper limit for the length u of a microsatellite in the second population. We assume that $T_0 = 3 \cdot 10^5$, $T = 5 \cdot 10^5$, $N_0 = N_1 = N_2 = 10^5$, $v_0 = v_1 = v_2 = 10^{-4}$ and $x = 12$.

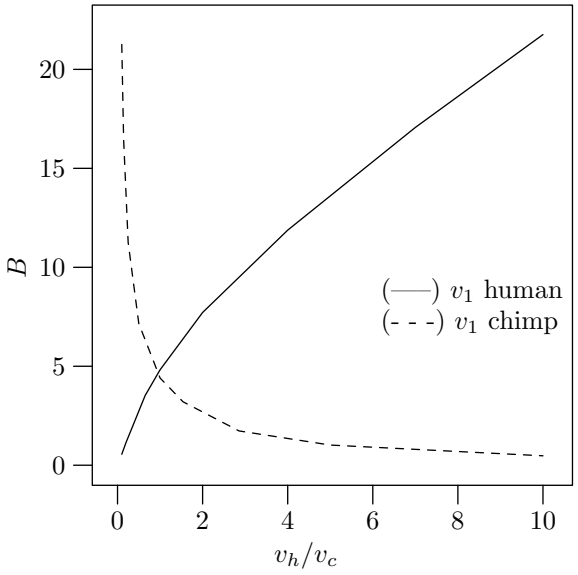

Figure 6.9: Reciprocal studies in the estimation of the ascertainment bias B between human and chimpanzee. We assume that $T_0 = 3 \cdot 10^5$, $T = 5 \cdot 10^5$ generations, $N_0 = 5 \cdot 10^4$, human effective population size is equal to $1 \cdot 10^4$, chimpanzee effective population size is equal to $3 \cdot 10^4$ and $v_0 = v_2 = 10^{-4}$ mutations per generation. Continuous line depicts results under the assumption that we choose a polymorphic locus from the human population and a dashed line presents the case when the polymorphic locus from the chimpanzee population is chosen.

crosatellite selected in one species is as likely to be longer as it is to be shorter than its homologue in the mother species. It seems also interesting to consider an upper boundary $u > x$ of the length of microsatellite above which the microsatellite experiences significant deletions or splits (177). Figure 6.8 presents results of a simulation of the upper limit of microsatellite length. For given demographic data, microsatellites longer than 30 tandem repeats occur so rarely that setting u for a value greater than 30 does not affect the value of the ascertainment bias.

7

Discussion

The subject of the dissertation are complex stochastic genetic systems. The background of this field of research is presented in Chapter 2. There we describe the main evolutionary forces and the basic methods used to model them. We also discuss the differences between backward-time and forward-time approaches. We continue this discussion in Chapter 3 by explaining the reasoning behind choosing between simulation and non-simulation methods. In this dissertation we focus on the non-simulation approaches as the aim of the dissertation concerns the useability of these methods and their refinement by specialized computer algorithms. Three different non-simulation stochastic genetic systems are introduced and examined in details in this dissertation (Chapters 4-6). The importance of each of these models along with the theses concerning them are explained in Chapter 3. The theses formulated in the dissertation are as follows:

- It is possible, using a non-simulation approach applied to the mathematical Moran model, to answer the question of the recombination identifiability, at least in the means of the relationships limited to a set of distributions, which jointly characterize allelic states at a number of different loci.

- It is possible, using a recursive algorithm, to calculate the exact distribution of the time to the MRCA of a large sample from a population evolved under any growth scenario with the time efficiency of the method allowing for analysis of large human populations.

- It is possible to build a non-simulation model of demographic interactions between many populations or species that can, in some applications, replace the

simulation-based approach.

With a constant development of analytical and simulation-based methods, increased interest in studies of recombination may be observed. The first model, described in Chapter 4, addresses some aspects of the dynamic behavior of recombination in a Moran model. More precisely, we answer the question of the asymptotic identifiability of a crossover recombination in the model with mutation and genetic drift. Similar analysis of the dynamics of the crossover recombination in the Moran model, although concerning other problems, may be found in (8, 9). Our model is a s-loci generalization of the two-loci model introduced in (18, 108). We explore the algorithms enabling construction of the transition probability matrices of the Markov chain describing the process (Section 4.4.2). Specialized hashing function based on the dynamic programming has been developed to ensure fast managing of the distributions. Our method works in $\mathrm{O}(s^4\varpi_s + \varpi_s^2)$ time complexity with very restrictive $20\varpi_s^2$[Bytes] memory complexity (Section 4.4.3), which is enough to obtain results for $s \leq 9$. Proper implementation of the sparse matrix operations would increase, with the cost of the time complexity, the feasible value of s. As a main result, we find that asymptotically the effects of recombination become indistinguishable from the effects of mutation and genetic drift (Theorem 4.1 and Section 4.3.1). This is very interesting, but rather paradoxical, result. However, one need to be aware that the result concerns asymptotic behavior only. Also, the framework of distributions used in the model is not complete. A set of distributions describing the relationships in the model jointly characterize allelic states at a number of loci at the same or different chromosome(s) but do not jointly characterize allelic states at a single locus on two or more chromosomes. As an example, probabilities such as $P[X_{11} = x_{11}; X_{12} = x_{12}; X_{23} = x_{23}]$ are included in the system, whereas probabilities such as $P[X_{11} = x_{11}; X_{21} = x_{21}; X_{23} = x_{23}]$ are not (Section 4.2.1). However, the system is sufficiently rich to allow computing all possible multipoint linkage disequilibria under recombination, mutation and drift, as well as their variances and covariances (Chapter 3 of (184)). Analysis of the Dobrušin's coefficient in the case $s = 3$ (Theorem 4.2) and the spectral gap theory in the general case (Section 4.5.2) suggest that the speed of convergence of the system is exponential. Comparison with the Hudson's Wright-Fisher coalescence model with recombination shows that the Moran model yields higher correlation of the time to the MRCA at two loci than the Wright-Fisher

model. We quantify this rather intuitive fact, although by the interpolation of the simulation results only (Section 4.5.3).

Chapter 5 contains description of the method that allows to calculate in a time efficient way the expected distribution of the time to the MRCA for sample of any size regardless of the population growth scenario $\{N_t\}$. The data required is population size over population history. The total time complexity of the algorithm is of the order of $\mathrm{O}(n^3 + n^2 T)$, where T is the number of discrete generations and n is the sample size (Section 5.3.1). Therefore, one may obtain the results in a short time even for $n \approx 10^3$ and the time period comparable to the time-span of modern humanity. Because of using the exact discrete time approach, our method differs from the usually applied continuous time diffusion approximation of the coalescence process. As an example, in Polański et al. (157) such a model has been used to estimate the population history based on the pairwise difference of individuals from the sample. Another interesting continuous time diffusion approach is a model used by Takahata (173, 175), who estimated, based on the coalescence process, the time to the MRCA in a constant size population under a complex selection model. The values of $g_{nk}(t)$, the probabilities that a sample of n alleles descended from exactly k distinct ancestral allelic lines t generations ago, are estimated in the Takahata's paper. Further on, these values are used in the analysis of the survivability of the ancestral alleles. The $g_{nk}(t)$ values are closely related to the α_{t_k} entries from our model (Expression (5.8) and Section 5.2). Our model is simple and fast enough to be successfully used instead of these methods as long as the sample size n is in the feasible range. Unlike a diffusion approximation, our model works well for a small population size. Based on (5.8), our method can provide additional useful data. The entries $\alpha_{t,k}$ of the matrix α, being the probabilities that the sample has exactly k ancestral lines at time t, give the exact distribution of the number of the ancestral lines over time. These values may be used to estimate population history based on the history of the sample(s) from that population (124) or to analyze the survivability of the ancient lineages in the population (175). The method, and the matrix α in particular, may also be used in studying the dynamics of the change of the MRCA over the time (155). As a result of the coalescence events between two generations, some of the genealogy lineages from the previous generation may not be continued in the following one. These deceased lineages may affect the time when the MRCA appears. To analyze this variability one needs to calculate and compare the distribution of the time to the MRCA for each

generation from a given period of time. As evident from Section 5.4.3, there exists a great variability of the estimated time to the MRCA distributions inferred from limited size samples. This variability is intrinsic in the statement of the problem in the sense that it reflects the wide mix of older and younger genealogies, which is affected by the history of bottlenecks and expansions. This variability is likely to affect the results of study of the age of founder mutations of genetic diseases and our method may be used to help to discriminate between various alternative architectures of genetic diseases based on population samples, as it was done in (153) using other methods. Studies of the time to the MRCA in the Poland and World populations (Section 5.4.2) lead to the apparently paradoxical finding that these times are longer than the time-span of modern humanity. However, it has to be noticed that this only means that fragments of the genome that did not recombine and were not under significant selection behave this way. In the case of the absence of complete demographic information over the whole period examined, the method can still be successfully applied provided that the major demographic events from the population history are incorporated in the growth scenario. The missing data between two consecutive modeled events can be interpolated, for example by exponential function, which should not cause a substantial impact on the results. The method can also be used as a testing platform to verify unknown demographic scenarios using genetic data. Besides the main method, Chapter 5 contains also the results of studies of the time to the MRCA of the population evolved according to the Galton-Watson process (Section 5.4.3). We obtain the Galton-Watson genealogies using our framework (Section 5.3.2) with a new algorithm that allows to store non-extinct genealogical lineages of the Galton-Watson process in a time and memory efficient way (Section 5.3.2). Finally, in Section 5.4.2 we present how to estimate an effective population size of real human population.

In Chapter 6 we present a model of demographic network that allow to calculate the joint probability distribution of a pair of individuals of different allelic types drawn from populations from the network. We model three types of discrete demographic events: splits, merges and migrations. Between these events, the network evolves according to the continuous-time coalescence model with mutation and genetic drift. Although the allelic space (given by the mutation model) does not change, the mutation intensities may vary between different populations or different time-intervals. We do not assume any population growth scenario. Evolution of the population is given by the Lyapunov

equation of the form of (6.10) but may be substituted by other continuous-time models (i.e., the model of the differences in the number of tandem repeats is used in Section 6.5.2). Number of optimizations, such as using of an adaptive step control algorithm in the Runge-Kutta method or a sparse matrix multiplication algorithm working in the square time-complexity, are used in the implementation of the model (Section 6.4.2). The complexity of the model strongly depends on the number of all possible entries of the joint distribution equal to $N_{\mathbb{A}}^n$, where $N_{\mathbb{A}}$ is the number of allelic types and n is a sample size (Section 6.4.3). Therefore, the model is usually used for $n = 2$ and $N_{\mathbb{A}} < 1000$. However, sometimes the model can be used for larger samples (Section 6.3.1) or larger allelic spaces (Section 6.3.2). The model does not include two important genetic forces: recombination or selection. Incorporating of the Hudson's coalescent recombination model (90) would lead to the change of Formula (6.10) into recursive equation with the joint distribution for a sample of size n dependent on the value of the distribution for a sample of size $n + 1$. However, our recombination model explained in Chapter 4 can be used in the demographic network model by incorporating Formula (4.12) into the model. Selection can added to the model in two ways (139) by using either the ancestral selection graphs (114) or the structured coalescent (102). We skip these refinements in the dissertation. Our model differs from the common simulation-based approaches. The obvious advantage of our method is that we obtain the exact results. One can calculate most of the genetic parameters that characterize populations and their interactions from the joint probability distribution of the allelic types. Despite its limits (especially the sample size and the length of the haplotype sequence that can be feasibly modeled), our method can be used in many real data application. The method works particularly well for the microsatellite mutation models (Sections 6.5.2 and 6.5.3). In Section 6.5 we present several example applications where our method can be useful. The main use of such a model lies in estimating of the parameter values under given demographic scenario. We estimate the equilibrium parameters (including the linkage disequilibrium) for a simple SNP model (Section 6.5.1), the pairwise difference in long haplotype sequences (Section 6.5.2), the Slatkin's distance between two populations (Section 6.5.2) and the ascertainment bias in a microsatellite model (Section 6.5.3). The results obtained from the latter example suggest higher microsatellite's mutation rate in human than in chimpanzee. Another interesting results concern the evolution of the microsatellite loci. We model the influence of the upper

limit of the number of tandem repeats on the microsatellite locus on the value of the ascertainment bias (177) and obtain that the limit set on the value greater than 30 does not affect the ascertainment bias. The estimates from the demographic model may be compared to the values obtained from genetic data and, therefore, can be used as measures in testing of the past demographic scenarios (170). We use such an approach to study of the common history of the Slavs and the Balts based on the Y-STR data and Slatkin's R_{ST} genetic distance (Section 6.5.2).

The results in this work can surely be a point of departure for further research. Our Moran model does not assume any particular mutation model. Thus, it may be interesting to study the dynamics of the model under specific mutation. The recombination scheme used in this model can also be used in other models (i.e., in the demographic network model). Model described in Chapter 5 can be used as a tool in many other models. We have already mentioned in the previous section of this chapter the most important applications, where this model can be useful (mainly as a testing platform to verify different demographic scenarios or disease architectures). Our demographic network model can be studied further on two levels: as a model or as a tool. On the field of modeling we may increase the useability of the model by: (i) introducing new genetic mechanisms to the model, (ii) optimizing performance of the model and (iii) preparing specialized algorithms for specific cases (such as a compression algorithm presented in Section 6.3.2). The real power of this model lies in using it as a tool in studies of the past demographies of different populations or species. In Section 6.5 we present several examples of such studies, but the real use of the model in this field is much wider.

Acknowledgements

I cannot imagine that this dissertation could be written without help from many people. I would like to thank all of them.

I would like to address special thanks to my supervisor, Professor Marek Kimmel, who motivated me all the time since our cooperation has begun. I am most grateful for his support, stimulating discussions, endless patience in explaining all difficulties and improving my works. This dissertation would be much worse without his remarks and suggestions. I am very grateful for inviting me, twice, for research internships in the Department of Statistics at the Rice University.

I would like to express my deep gratitude to Professor Adam Bobrowski, who introduced me to the field of population genetics and bioinformatics. All of our short, but very constructive discussions allowed me to set first steps in this, new for me, field of science and confirmed that my choice of research directions is correct.

Significant part of this dissertation treats about constructing of the computer algorithms. My interest in this field has been greatly stimulated by participating in many polish and international programming contests. Here, I want to thank all people who competed or organized these contests. I would like to express special thanks to Marcin Ciura who was my teammate and coach, and to Jacek Widuch who organized Liga Zadaniowa and introduced me into the field of algorithms. I want also to thank all other people who were my teammates or helped me to improve my knowledge about algorithms: Przemysław Drochomirecki, Dariusz Czechowski, Marcin Ćwięk and Damian Szczepanik.

It is true that many solutions occur when one can look at the problem from different points of view. There is no better way to do so than talk about

it with other people. Thus, I would like to acknowledge Łukasz Olczak, Dariusz Myszor, Doctor Krzysztof Cyran and Professor William Amos who shared with me and explained me their ideas. I wish I had more time to study them in details. Also, I would like to thank Professor Andrzej Polański for allowing me to participate in his seminars and for sharing his opinions on my work.

During my Ph.D. study I was employed in the Software Department of the Silesian University of Technology. I am thankful for the possibility to work in this team. Many thanks go to the manager Doctor Przemysław Szmal, who motivated me in my work. I am not able to enumerate all the staff, so I would like to give special thanks to my closest colleagues: Jacek Widuch, Marcin Ciura, Michał Świderski, Krzysztof Simiński, Adam Karwan and Tomasz Gandor.

Finally, an exceptional 'Thank you!' I would like to say to my family who gave me best possible upbringing, and encouraged and supported me in my decisions.

References

[1] **Human Genome Project**. `http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml`. 1

[2] **International HapMap Project**. `http://hapmap.ncbi.nlm.nih.gov/`. 1

[3] **International HapMap Project**. `http://www.1000genomes.org/`. 2

[4] **Wolfram Mathematica**. `http://www.wolfram.com/mathematica/`. 36

[5] Milton Abramowitz and Irene A. Stegun. *Integration, chapter 25.4 in: Handbook of mathematical functions with formulas, graphs, and mathematical tles, 9th printing*. Courier Dover Publications, New York, 1972. 81

[6] Krishna B. Athreya and Peter E. Ney. *Branching Processes*. Springer, New York, 1972. 12

[7] Adam Auton and Gil McVean. **Recombination rate estimation in the presence of hotspots**. *Genome Research*, **17(8)**:1219–1227, 2007. 9

[8] Ellen Baake and Inke Herms. **Single-crossover dynamics: finite versus infinite populations**. *Bulletin of Mathematical Biology*, **70(2)**:603–624, 2008. 22, 98

[9] Ellen Baake and Thiemo Hustedt. **Moment closure in a Moran model with recombination**. *arXiv:1105.0793v1 [math.PR]*, 2011. 98

[10] Maximilian O. Baldia. **The Corded Ware / Single Grave Culture**. `http://www.comp-archaeology.org/CordedWare.htm`, 2006. 88

[11] Francois Balloux. **EASYPOP (Version 1.7): A computer program for population genetics simulations**. *Journal of Heredity*, **92(3)**:301–302, 2001. 14, 22

[12] Peter Beerli. **Comparison of Bayesian and maximum-likelihood inference of population genetic parameters**. *Bioinformatics*, **22(3)**:341–345, 2006. 16

[13] Elise M. S. Belle et al. **Comparing models on the genealogical relationships among Neandertal, Cro-Magnoid and modern Europeans by serial coalescent simulations**. *Heredity*, **102**:218–225, 2009. 86, 87

[14] Elise M. S. Belle, Uma Ramakrishnan, Joanna L. Mountain, and Guido Barbujani. **Serial coalescent simulations suggest a weak genealogical relationship between Etruscans and modern Tuscans**. *Proceedings of the National Academy of Sciences of the United States of America*, **103**:8012–8017, 2006. 86

[15] Richard Bellman. **The theory of dynamic programming**. *Bulletin of the American Mathematical Society*, **60(6)**:503–515, 1954. 3, 56

[16] Richard Bellman and Theodore E. Harris. **On the theory of age-dependent stochastic branching processes**. *Proceedings of the National Academy of Science of the United States of America*, **34(12)**:601–604, 1948. 12

[17] Adam Bobrowski. *Functional analysis for probability and stochastic processes*. Cambridge University Press, 2005. 53

[18] Adam Bobrowski and Marek Kimmel. **A random evolution related to a Fisher-Wright-Moran model with mutation, recombination and drift**. *Mathematical Methods in the Applied Sciences*, **26**:1587–1599, 2003. 28, 29, 34, 52, 98

[19] Adam Bobrowski, Marek Kimmel, Ovide Arino, and Ranjit Chakraborty. **A semigroup representation and asymetric behavior of certain statistics of the Fisher-Wright-Moran coalescent**. *Handbook of Statistics*, **19**:215–242, 2001. 77

[20] Andrzej Borzyskowski. **The Slavic Ethnogenesis**. `http://www.andrzejb.net/slavic/`. 88

[21] Rodney E. Canfield and Carl Pomerance. **On the problem of uniqueness for the maximum Stirling number(s) of the second kind**. *Electronic Journal of Combinatorial Number Theory*, **2(2002), Paper A01 electronic only**:13–13, 2002. 56

[22] Jeff R. Cash and Alan H. Karp. **A variable order Runge-Kutta method for initial value problems with rapidly varying right-hand sides**. *ACM Transactions on Mathematical Software*, **16(3)**:201–222, 1990. 83

[23] Stanislaw Cebrat and Dietrich Stauffer. **Altruism and antagonistic pleiotropy in Penna ageing model**. *Theory in Biosciences*, **123(3)**:235–241, 2005. 10

[24] Mark Collard and Nickolas Franchino. **Pairwise difference analysis in modern human origins research**. *Journal of Human Evolution*, **43(3)**:323–352, 2002. 24

[25] International HapMap Consortium. **The International HapMap Project**. *Nature*, **426(6968)**:789–796, 2003. 1

[26] The 1000 Genomes Project Consortium. **A map of human genome variation from population-scale sequencing**. *Nature*, **467(7319)**:1061–1073, 2010. 2

[27] Graham Coop and Robert C. Griffiths. **Ancestral inference on gene trees under selection**. *Theoretical Population Biology*, **66(3)**:219–232, 2004. 14

[28] Gillian Cooper, David C. Rubinsztein, and William Amos. **Ascertainment bias cannot entirely account for human microsatellites being longer than their chimpanzee homologues**. *Human Molecular Genetics*, **7(9)**:1425–1429, 1998. 92, 94

[29] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms (2nd edition)*. MIT Press and McGraw-Hill Book Company, 2001. 3, 56

[30] Allan M. Crawford et al. **Microsatellite evolution: testing the ascertainment bias hypothesis**. *Journal of Molecular Evolution*, **46**:256–260, 1998. 91

[31] James F. Crow. **Hardy, Weinberg and language impediments**. *Genetics*, **152(3)**:821–825, 1999. 1, 8

[32] Charles Darwin. *On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life*. John Murray, 1859. 1

[33] Charles Darwin and Alfred R. Wallace. **On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection**. *Linnean Society of London*, **Zoology 3**:46–50, 1858. 1

[34] Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976. 10

[35] Bruno de Finetti. **Sul Significato Soggettivo della Probabilitia**. *Fundamenta Mathematicae*, **17**:298–329, 1931. 78

[36] Gerda de Vries et al. *A Course in Mathematical Biology: Quantitative Modeling with Mathematical and Computational*. SIAM, 2006. 16

[37] Roland Dobrušin. **The central limit theorem for nonhomogeneous Markov chains**. *Verojatnost. i Primenen*, **1**:365–425, 1956. 27

[38] Richard Durret. *Probability Models for DNA Sequence Evolution*. Springer, New York, 2002. 9, 22

[39] Albert Einstein. **On the theory of the brownian movement**. *Annals of Physics*, **19**:371–381, 1906. 18

[40] Hans Ellegren. **Microsatellites: simple sequences with complex evolution**. *Nature Reviews Genetics*, **5**:435–445, 2004. 7

[41] Hans Ellegren et al. **Microsatellite evolution–a reciprocal study of repeat lengths at homologous loci in cattle and sheep**. *Molecular Biology and Evolution*, **14(8)**:854–860, 1997. 92

[42] Klaus J. Engel and Rainer Nagel. *One-parameter semigroups for linear evolution equations*. Springer, Berlin, 2000. 25

[43] Steven N. Evans and Peter L. Ralph. **Dynamics of the time to the most recent common ancestor in a large branching population**. *Annals of Applied Probability*, **20(1)**:1–25, 2010. 23

[44] Warren J. Ewans. *Mathematical Population Genetics*. Springer, Berlin, 2004. 2, 10, 18

[45] Warren Ewens. *Mathemtical Population Genetics*. Springer-Verlag, New York, 2004. 10

[46] Laurent Excoffier, Guillaume Laval, and Stefan Schneider. **Arlequin, An intergrated software package for population genetics data analysis**. `http://http://cmpg.unibe.ch/software/arlequin3`. 89

[47] Marcus W. Feldman, Aviv Bergman, David D. Pollock, and David B. Goldstein. **Microsatellite genetic distances with range constraints: analytic description and problems of estimation**. *Genetics Society of America*, **145(1)**:207–216, 1997. 8

[48] Marcus W. Feldman and Jonathan K. Pritchard. **Statistics for microsatellite variation based on coalescence**. *Theoretical Population Biology*, **50**:325–344, 1996. 48

[49] William Feller. *An introduction to probability theory and its applications, vol. 2*. Wiley, 1971. 18

[50] Joseph Felsenstein. **Accuracy of coalescent likelihood estimates: Do we need more sites, more sequences, or more loci?** *Molecular Biology and Evolution*, **23**:691–700, 2005. 16

[51] Ronald A. Fisher. **On the dominance ratio**. *Proceedings of the Royal Society of Edinburgh*, **42**:321–341, 1922. 12

[52] Ronald A. Fisher. *The genetical theory of natural selection*. Clarendon Press, Oxford, 1930. 1, 10, 11

[53] George E. Forsythe, Michael A. Malcolm, and Cleve B. Moler. *Computer methods for mathematical computations*. Prentice-Hall, 1977. 83

[54] Zoran Gajic, Muhammed Tahir, and Javed Qureshi. *Lyapunov matrix equation in system stability and control*. Academic Press, San Diego, 1995. 78

[55] William C. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, 1971. 83

[56] Stuart Geman and Donald Geman. **Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6(6)**:721–741, 1984. 18

[57] Aleeza C. Gerstein et al. **Haploids adapt faster than diploids across a range of environments**. *Journal of Evolutionary Biology*, **24(3)**:531–540, 2011. 65

[58] Walter G. Gilks et al. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, 1996. 17

[59] John H. Gillespie. *Population genetics: a concise guide, 2nd edition*. The Johns Hopkins University Press, 2004. 90

[60] Gene H. Golub and Charles F. Van Loan. *Matrix computations, 3rd edition*. Johns Hopkins University Press, Baltimore, 1996. 38

[61] Simon J. Goodman. **RST Calc: a collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance**. *Molecular Ecology*, **6**:881–885, 1997. 89

[62] Boris V. Gornung. *Iz predystorii obrazovaniia obshcheslavianskogo iazykovogo edinstva*. Izd-vo Akademii nauk SSSR, Moskva, 1963. 88

[63] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete mathematics : a foundation for computer science*. Addison-Wesley, 1994. 30, 31, 53

[64] Richard Griego and Reuben Hersh. **Theory of random evolutions with applications to partial differential equations**. *Transactions of the American Mathematical Society*, **156**:405–418, 1971. 34

[65] Robert C. Griffiths. **Lines of descent in the diffusion approximation of neutral Fisher-Wright models**. *Theoretical Population Biology*, **17**:37–50, 1980. 23

[66] Robert C. Griffiths. **Neutral two-locus multiple allele models with recombination**. *Theoretical Population Biology*, **19(2)**:169–186, 1981. 9, 14, 22, 48

[67] Robert C. Griffiths and Simon Tavare. **Sampling theory in neutral alleles in a varying environment**. *Philosophical Transactions: Biological Sciences*, **344(1310)**:403–410, 1994. 51

[68] Robert C. Griffiths and Simon Tavare. **Monte Carlo inference methods in population genetics**. *Mathematical and Computer Modelling*, **23**:141–158, 1996. 17

[69] Eupedia group. **Origins, age, spread and ethnic association of European haplogroups and subclades**. `http://www.eupedia.com/europe/origins_haplogroups_europe.shtml`. 88

[70] Nostratic group. **Indo-European chronology**. `http://indoeuro.bizland.com/project/chron/chron1.html`. 88

[71] Frederic Guillaume and Jacques Raugemont. **Nemo: an evolutionary and popuation genetics programming framework**. *Bioinformatics*, **22(20)**:2556–2557, 2006. 22

[72] Silvia Guimaraes et al. **Genealogical Discontinuities among Etruscan, Medieval, and Contemporary Tuscans**. *Molecular Biology abd Evolution*, **26(9)**:2157–2166, 2009. 24

[73] John B. S. Haldane. **A mathematical theory of natural and artificial selection**. *Mathematical Proceedings of the Cambbridge Philosophical Society*, **23**:838–844, 1927. 12

[74] Matthew B. Hamilton. *Population Genetics*. Wiley, 2009. 15

[75] William D. Hamilton. **The evolution of altruistic behavior**. *American Naturalist*, **97**:354–356, 1963. 10

[76] William D. Hamilton. **The genetical evolution of social behaviour**. *Journal of Theoretical Biology*, **7(1)**:1–52, 1964. 10

[77] Kenneth M. Hanson and Gregory S. Cunningham. **Posterior sampling with improved efficiency**. *Proceedings of SPIE, Medical Imaging: Image Processing*, **3338**:371–382, 1998. 17

[78] Theodore E. Harris. *The Theory of Branching Processes*. Springer, Berlin, 1963. 12

[79] W. Keith Hastings. **Monte Carlo Sampling Methods Using Markov Chains and Their Applications**. *Biometrika*, **57(1)**:97–109, 1970. 18

[80] Jotun Hein, Mikkel H. Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, 2005. 7, 9

[81] Ines Hellmann et al. **A neutral explanation for the correlation of diversity with recombination rates in humans**. *American Journal of Human Genetics*, **72(6)**:1527–1535, 2003. 6

[82] Paul L. Hennequin and Albert Tortrat. *Probability Theory and Some of its Applications*. Masson et Cie, Paris, 1965. 26

[83] Jody Hey. **FPG: a computer program for forward population genetic simulation**. `http://lifesci.rutgers.edu/~heylab/HeyLabSoftware.htm`. 22

[84] Jody Hey. **What's so hot about recombination hotspots?** *PLoS Biology*, **2(6)**:e190. doi:10.1371/journal.pbio.0020190, 2004. 8

[85] Asger Hobolth, Ole F. Christensen, Thomas Mailund, and Mikkel H. Schierup. **Genetic relationships and speciation times of human, chimpanzee and gorilla inferred from a coalescent hidden Markov model**. *PLoS Genetics*, **3(2)**:e7. doi:10.1371/journal.pgen.0030007, 2007. 93, 94

[86] Asger Hobolth et al. **Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection**. *Genome Research*, **2011(21)**:349–356, 2011. 94

[87] Clive J. Hoggart et al. **FREGENE: software for simulating large genomic regions**. `http://www.ebi.ac.uk/projects/BARGEN/download/FREGEN/`. 22

[88] Susan L. Holbeck and Jeffrey N. Strathern. **A role for REV3 in mutagenesis during double-strand break repair in Saccharomyces cerevisiae**. *Genetics*, **147**:1017–1024, 1997. 6

[89] Neil Howell, Iwona Kubacka, and David A. Mackey. **How rapidly does the human mitochondrial genome evolve?** *American Journal of Human Genetics*, **59(3)**:501–509, 1996. 5

[90] Richard R. Hudson. **Properties of a neutral allele model with intragenic recombination**. *Theoretical Population Biology*, **23(2)**:183–201, 1983. 1, 3, 9, 14, 15, 22, 38, 45, 48, 101

[91] Richard R. Hudson. **Gene genealogies and the coalescent process**. *Douglas J. Futuyma, Janis Anotonovics (eds) Oxford Surveys in Evolutionary Biology*, **7**:1–44, 1990. 3, 45

[92] Richard R. Hudson. **Two-locus sampling distributions and their application**. *Genetics*, **159**:1805–1817, 2001. 9

[93] Richard R. Hudson. **Generating samples under a Wright-Fisher neutral model of genetic variation**. *Bioinformatics*, **18(2)**:337–338, 2002. 45

[94] Berlin Institute of Legal Medicine, Humboldt-University. **European Y-chromosome microsatellite data**. `http://www.ystr.org`. 89

[95] Entire issue of Nature. **The human genome**. *Nature*, **409(6822)**:745–964, 2001. 1

[96] Entire issue of Science. **The human genome**. *Science*, **291(5507)**:1145–1434, 2001. 1

[97] Peter Jagers. *Branching Processes with Biological Applications*. Wiley, London, 1975. 12

[98] Philippe Jarne and Pierre J. L. Lagoda. **Microsatellites, from molecules to populations and back**. *Trends in Ecology and Evolution*, **11(10)**:424–429, 1996. 7

[99] Elena Jazin et al. **Mitochondrial mutation rate revisited: hot spots and polymorphism**. *Nature Genetics*, **18**:109–110, 1998. 5

[100] Andrzej Jezierski and Cecylia Leszczyńska. *Historia gospodarcza Polski*. Key Text Wydawnictwo, 2003. 125

[101] Lynn B. Jorde. *The genetic structure of subdivided human populations: a review; in: Current developments in anthropological genetics. Volume 1: Theory and methods (James H. Mielke and Michael H. Crawford, eds)*. Plenum Press, New York, 1980. 65, 88

[102] Norman L. Kaplan, Thomas Darden, and Richard R. Hudson. **The coalescent process in models with selection**. *Genetics*, **120**:819–829, 1988. 101

[103] Manfred Kayser et al. **Evaluation of Y-chromosomal STRs: a multicenter study**. *International Journal of Legal Medicine*, **110**:125–129, 1997. 89

[104] Manfred Kayser, Edward J. Vowles, Dennis Kappei, and William Amos. **Microsatellite length differences between humans and chimpanzees at autosomal loci are not found at equivalent haploid Y chromosomal loci**. *Genetics*, **173(4)**:2179–2186, 2006. 92

[105] Marek Kimmel and David E. Axelrod. *Branching Processes in Biology*. Springer Varlag, New York, 2002. 12, 13, 17

[106] Marek Kimmel and Ranajit Chakraborty. **Measures of variation at DNA repeat loci under a general stepwise mutation model**. *Theoretical Population Biology*, **50(3)**:345–367, 1996. 7, 8

[107] Marek Kimmel et al. **Signatures of population expansion in microsatellite repeat data**. *Genetics*, **148**:1921–1930, 1998. 90

[108] Marek Kimmel and Joanna Polańska. **A model of dynamics of mutation, genetic drift and recombination in DNA-repeat genetic loci**. *Archives of Control Sciences*, **9(XVL)**:143–157, 1999. 28, 98

[109] Motoo Kimura. **A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences**. *Journal of Molecular Evolution*, **16(2)**:111–120, 1980. 7

[110] John F. C. Kingman. **The coalescent**. *Stochastic Processes and their Applications*, **13(3)**:235–248, 1982. 1, 14, 16, 23

[111] Jon Kleinberg and Eva Tardos. *Algorithm designs*. Addison-Wesley, 2006. 56

[112] Tadeusz Korzon. *Wewnętrzne dzieje Polski za Stanisława Augusta*. Warszawa, 1882. 125

[113] Michael Kremer. **Population Growth and Technological Change: One Million B.C. to 1990**. *The Quarterly Journal of Economics*, **108(3)**:681–716, 1993. 64, 88, 124

[114] Stephen M. Krone and Claudia Neuhauser. **Ancestral processes with selection**. *Theoretical Population Biology*, **51(3)**:210–237, 1997. 14, 101

[115] Mary K. Kuhner. **LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters**. *Bioinformatics*, **22(6)**:768–770, 2006. 16

[116] Cezary Kulko and Andrzej Wyczański. *Historia Polski w liczbach. Ludność. Terytorium*. Zakład Wydawnictw Statystycznych, Warszawa, 1993. 125

[117] Tadeusz Lalik. *Encyklopedia historii gospodarczej Polski do 1945*. Warszawa, 1981. 125

[118] Alan J. Laub. *Matrix analysis for scientists and engineers*. SIAM: Society for Industrial and Applied Mathematics, 2005. 79

[119] Richard C. Lewontin. **The interaction of selection and linkage. General considerations; heterotic models**. *Genetics*, **49(1)**:49–67, 1964. 85

[120] Ray A. Littler. **Loss of variability at one locus in a finite population**. *Mathematical Biosciences*, **25**:151–163, 1975. 23

[121] Jerzy Lukowski and Hubert Zawadzki. *A concise history of Poland*. Cambridge University Press, 2001. 125

[122] Andrey Markov and Ronald A. Howard. **Extension of the limit theorems of probability theory to a sum of variables connected in a chain**. *Dynamic Probabilistic Systems*, **1: Markov Models**:552–577, 1971. 16

[123] George Marsaglia, Arif Zaman, and Wai Wan Tsang. **Toward a universal random number generator**. *Letters in Statistics and Probability*, **9(1)**:35–39, 1990. 59

[124] Yosef E. Maruvka, Nadav M. Shnerb, Yaneer Bar-Yam, and John Wakeley. **Recovering population parameters from a single gene genealogy: an unbiased estimator of the growth rate**. *Molecular Biology and Evolution, In Press*, **28(5)**:1617–1631, 2011. 55, 99

[125] Gil McVean, Philip Awadalla, and Paul Fearnhead. **A coalescent-based method for detecting and estimating recombination from gene sequences**. *Genetics*, **160**:1231–1241, 2002. 9, 16

[126] Gil McVean et al. **The fine-scale structure of recombination rate variation in the human genome**. *Science*, **304(5670)**:581–584, 2004. 9

[127] Gregor J. Mendel. **Versuche uber Pflanzen-Hybriden**. *Verhandlungen des naturforschenden Vereines in Brűnn*, **IV(1865)**:3–47, 1865. 1

[128] Nicholas Metropolis. **The beginning of the Monte Carlo method**. *Los Alamos Science*, **15**:125–130, 1987. 1, 17

[129] Nicholas Metropolis and Stanisław Ulam. **The Monte Carlo method**. *Journal of the American Statistical Association*, **44(247)**:335–341, 1949. 1, 17

[130] Nicholas Metropolis et al. **Equations of state calculations by fast computing machines**. *Journal of Chemical Physics*, **21(6)**:1087–1092, 1953. 18

[131] Mariusz Milik and Jeffrey Skolnick. **Insertion of peptide chains into lipid membranes: an off-lattice Monte Carlo dynamics model**. *Proteins*, **15**:10–25, 1993. 17

[132] P. A. P. Moran. **A general theory of the distribution of gene frequencies**. *Proceedings of the Royal Society, B Biological Sciences*, **149(934)**:113–116, 1958. 11

[133] Jose E. Moyal. **The general theory of stochastic population processes**. *Acta Mathematica*, **108(1)**:1–31, 1961. 17

[134] Chen Mufa. **Estimation of spectral gap for Markov chains**. *Acta Mathematica Sinica*, **12(4)**:337–360, 1996. 28

[135] Michael W. Nachman and Susan L. Crowell. **Estimate of the mutation rate per nucleotide in humans**. *Genetics*, **156**:297–304, 2000. 5

[136] United Nations. **The World at Six Billion Project**. `http://www.un.org/esa/population/publications/sixbillion/sixbilpart1.pdf`, 2000. 124

[137] Maarten J. Nauta and Franz J. Weissing. **Constraints on allele size at microsatellite loci: implications for genetic differentiation**. *Genetics Society of America*, **143(2)**:1021–1032, 1996. 8

[138] Rasmus Nielsen and Per J. Palsbøll. **Single-locus tests of microsatellite evolution: multi-step mutations and constraints on allele size**. *Molecular Phylogenetics and Evolution*, **11(3)**:477–484, 1999. 7

[139] Magnus Nordborg. *Coalescent Theory. In: Handbook of Statistical Genetics, David J. Balding, Martin Bishop, Chris Cannings eds.* John Wiley and Sons, Chichester, 2001. 101

[140] Leonard Nunney. **The influence of mating system and overlapping generations on effective population size**. *Evolution*, **47(5)**:1329–1341, 1993. 65, 88

[141] Neil O'Connell. **The genealogy of branching processes and the age of our most common ancestor**. *Advances Applied Probability*, **27(2)**:418–442, 1995. 13

[142] Central Statistical Office of Poland. **Struktura ludności Polski**. `http://www.stat.gov.pl/cps/rde/xbcr/gus/PUBL_lu_struktura_ludnosci_02_tablica2.xls`, 2009. 125

[143] Brendan O'Fallon. **TreesimJ: a flexible, forward time population genetic simulator**. *Bioinformatics*, **26**:2200–2201, 2010. 14, 22

[144] Tomoko Ohta and Motoo Kimura. **A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population**. *Genetical Research*, **22(2)**:201–204, 1973. 7, 17

[145] Henryk Łowmiański. *Polityka Jagiellonów*. Wydawnictwo Poznańskie, 1999. 125

[146] Badri Padhukasahasram et al. **Exploring population genetic models with recombination using efficient forward-time simulations**. *Genetics*, **178(4)**:2417–2427, 2008. 22

[147] Athanasios Papoulis. *Probability, random variables, and stochastic processes, 2nd edition*. McGrow-Hill, New York, 1984. 16, 18

[148] Thomas J. Parsons and Mitchell M. Holland. **Replay to Jazin et al.: Mitochondrial mutation rate revisited: hot spots and polymorphism**. *Nature Genetics*, **18**:110, 1998. 5

[149] Jonathan R. Partington. *Linear operators and linear systems*. London Mathematical Society Student Texts, Cambridge University Press, 2004. 26

[150] Amnon Pazy. *Semigroups of linear operators and applications to partial differential equations*. Springer, New York, 1983. 78

[151] Bo Peng, Christopher I. Amos, and Marek Kimmel. **Forward-time simulations of complex human diseaes**. *PLoS Genetics*, **3(3)**:e47. doi:10.1371/journal.pgen.0030047, 2007. 14

[152] Bo Peng and Marek Kimmel. **simuPop: A forward-time population genetics simulation environment**. *Bioinformatics*, **21**:3686–3687, 2005. 9, 14, 22

[153] Bo Peng and Marek Kimmel. **Simulations provide support for the common disease common variant hypothesis**. *Genetics*, **175**:763–776, 2007. 100

[154] Anna Perez-Lezaun et al. **Population Genetics of T-chromosome short tandem repeats in humans**. *Journal of Molecular Evolution*, **45(3)**:265–270, 1997. 88

[155] Peter Pfaffelhuber and Anton Wakolbinger. **The process of most recent common ancestors in an evolving coalescent**. *Stochastic Processes and their Applications*, **116(12)**:1836–1859, 2006. 23, 99

[156] RafałPloski et al. **Homogeneity and distinctiveness of Polish paternal lineages revealed Y chromosome microsatellite haplotype analysis**. *Human Genetics*, **110(2002)**:592–600, 2002. 88

[157] Andrzej Polański, Marek Kimmel, and Ranajit Chakraborty. **Application of a time-dependent coalescence process for inferring the history of population size changes from DNA sequence data**. *Proceedings of the*

National Academy of Science of the United States of America, **95**:5456–5461, 1998. 99

[158] William H. Press, Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling. *Numerical recipes in C - the art of scientific computing*. Cambridge University Press, 1988. 83

[159] PWN. *Encyklopedia PWN*. Wydawnictwo Naukowe PWN, 2006. 125

[160] Paul Ressel. **De Finetti-type theorems: an analytical approach**. *The Annals of Probability*, **13(3)**:898–922, 1985. 78

[161] Mark Ridley. *Evolution, Third Edition*. Wiley-Blackwell, 2004. 8

[162] Anna Di Rienzo and Richard R. Hudson. **An evolutionary framework for common diseases: the ancestral-susceptibility model**. *Trends in Genetics*, **21(11)**:596–601, 2005. 14

[163] Ryszard Rudnicki et al. **Markov semigroups and their applications**. *Lecture Notes in Physics*, **597**:215–238, 2002. 25

[164] Raazesh Sainudiin, Andrew G. Clark, and Richard T. Durrett. **Simple models of genomic variation in human SNP density**. *BMC Genomics*, **8:146**:doi:10.1186/1471–2164–8–146, 2007. 9

[165] Laurent Saloff-Coste. **Lectures on finite Markov chains; in: Lectures on probability theory and statistics**. *Lecture Notes in Mathematics*, **1665**:301–413, 1997. 28

[166] Shlomo Sawilowsky. **You think you've got trivials?** *Journal of Modern Applied Statistical Methods*, **2(1)**:218–225, 2003. 17

[167] Damien Simon and Bernard Derrida. **Evolution of the most recent common ancestor of a population with no selection**. *Journal of Statistical Mechanics*, **(2006) P05002**:10.1088/1742–5468/2006/05/P05002, 2006. 23

[168] Montgomery Slatkin. **A measure of population subdivision based on microsatellite allele frequencies**. *Genetics*, **139**:457–462, 1995. 89

[169] Chris C. A. Spencer et al. **The influence of recombination on human genetic diversity**. *PLoS Genetics*, **2(9)**:e148. doi:10.1371/journal.pgen.0020148, 2006. 9

[170] Mark Stoneking and Johannes Krause. **Learning about human population history from ancient and modern genomes**. *Nature Reviews Genetics*, **12**:603–614, 2011. 85, 102

[171] Fumio Tajima. **Evolutionary relationship of DNA sequences in finite populations**. *Genetics*, **105(2)**:437–460, 1983. 1

[172] Fumio Tajima. **Statistical methods for testing the neutral mutation hypothesis by DNA polymorphism**. *Genetics*, **123(3)**:585–595, 1989. 6

[173] Naoyuki Takahata. **A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism**. *Proceedings of the National Academy of Science of the United States of America*, **87**:2419–2423, 1990. 99

[174] Naoyuki Takahata. **Allelic genealogy and human evolution**. *Molecular Biology and Evolution*, **10(1)**:2–22, 1993. 94

[175] Naoyuki Takahata. *Evolutionary Genetics of Human Paleo-Populations. In: Mechanisms of Molecular Evolution, Naoyuki Takahata and Andrew G. Clark eds.* Japan Scietific Societies Press, Tokio, 1993. 99

[176] Christoph W. Ueberhuber. *Numerical Computation 2: methods, software and analysis*. Springer-Verlag, Berlin, 1997. 81

[177] Edward J. Vowels and William Amos. **Quantifying ascertainment bias and species-specific length differences in human and chimpanzee microsatellites using genome sequences**. *Molocular Biology Evolution*, **23(3)**:598–607, 2006. 24, 92, 96, 102

[178] John Wakeley. **The limits of theoretical population genetics**. *Genetics Society of America*, **169(1)**:1–7, 2005. 14

[179] John Wakeley. *Coalescent Theory: An Introduction*. Ben Roberts Rublishing, 2008. 9, 11, 22, 23

[180] Henry W. Watson and Francis Galton. **On the probability of the extinction of families**. *Journal of the Anthropological Institute of Great Britain*, **4**:138–144, 1874. 12

[181] Geoffrey A. Watterson. **On the number of segregation sites in genetical models without recombination**. *Theoretical Population Biology*, **7(2)**:256–276, 1975. 6, 52

[182] Matthew T. Webster, Nick G. C. Smith, and Hans Ellegren. **Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments**. *Proceedings of the National Academy of Sciences of the United States of America*, **99(13)**:8748–8753, 2002. 94

[183] Stefan Weinzierl. **Introduction to Monte Carlo methods**. 2000. 17

[184] Bruce S. Weir. *Genetic data analysis II: methods for discrete population genetic data*. Sinauer Associates Inc, 1996. 98

[185] Eric Weisstein. **Chapman-Kolmogorov Equation**. `http://mathworld.wolfram.com/Chapman-KolmogorovEquation.html`. 16

[186] Carsten Wiuf. *Highly Structured Stochastic Systems, chapter 14*. Oxford University Press, 2003. 23

[187] Sewall Wright. **Evolution in Mendelian populations**. *Genetics*, **16(2)**:97–159, 1931. 1, 11

[188] Lev A. Zhivotovsky. **A new genetic distance with application to constrained variation at microsatellite loci**. *Molecular Biology and Evolution*, **16(4)**:467–471, 1999. 8

Appendix A

Population size of Poland and of the World

A. POPULATION SIZE OF POLAND AND OF THE WORLD

Table A.1: Population size of the World

year	size [10^6]	source
-1000000	0.125	Reference (113)
-10000	4	Reference (113)
-5000	5	Reference (113)
-4000	7	Reference (113)
-3000	14	Reference (113)
-2000	27	Reference (113)
-1000	50	Reference (113)
-500	100	Reference (113)
-200	150	Reference (113)
1	170	Reference (113)
200	190	Reference (113)
400	190	Reference (113)
600	200	Reference (113)
800	220	Reference (113)
1000	265	Reference (113)
1100	320	Reference (113)
1200	360	Reference (113)
1300	360	Reference (113)
1400	350	Reference (113)
1500	425	Reference (113)
1600	545	Reference (113)
1650	545	Reference (113)
1700	610	Reference (113)
1750	720	References (113, 136)
1800	900	References (113, 136)
1850	1200	References (113, 136)
1875	1325	References (113, 136)
1900	1625	References (113, 136)
1950	2519	References (113, 136)
1975	4068	References (113, 136)
2000	6070	Reference (136)

Table A.2: Population size of Poland

year	size [10^6]	source
1000	1	Reference (121)
1375	1.9	Reference (116)
1400	4.2	Reference (145)
1575	7.5	Reference (117)
1650	11	Reference (117)
1750	12	Reference (100)
1775	12	Reference (100)
1800	9	Reference (112)
1850	11.1	Estimates available on the Internet based on the census in Prussia in 1846 and censuses from the Kingdom of Poland (Russia territory) and Galicja in 1870s-1890s
1900	20	Reference (159)
1925	35	Reference (159)
1950	25	Reference (142)
1975	34	Reference (142)
2000	38.2	Polish Demographic Yearbook 2000

List of Algorithms

List of Figures

List of Tables