Data Mining in Molecular Biology: A Journey From Raw Datasets to Biologically Meaningful Conclusions

Roman Jaksik, *Silesian University of Technology*

Abstract

21st century computer science allows to reach a new complexity level of computational analysis which is still rapidly growing beyond our imagination, constantly increasing the limits of human possibilities. But despite the technological advancement the limit is set by the human mind which is not capable of grasping such huge amounts of information. The problem of modern bioinformatics doesn't always concern the tremendous calculations time but the amount of experiment results that we produce which sometimes require complex processing to be acquired by a human. This work presents the possibilities of modern genomics available thanks to the data and analysis methods, gathered trough the last quarter of a century and problems which arise when one aims to test hundreds of independent hypothesis in order to reveal the secrets of the DNA.

1. Introduction

Today's technology allows to generate hundreds of gigabytes of DNA and RNA sequencing data in just few days for an incredibly low coast comparing to just a decade ago. The enormous rate of data generation by the low-cost, high-throughput technologies in genomics makes in most of the cases data processing a much harder and more time consuming task than the data gathering itself.

Such large-scale and high-dimensional data sets are impossible to be interpret without the use of appropriate bioinformatic methods and modern computer technology, but despite our incredibly fast growing advancement in this field there is still a big doubt if we are able to properly interpret the information hidden in the data sets, or did the data gathering methods outrace our data mining algorithms. This is proven by our continuously increasing understanding of the data provided by the Human Genome Project which ended over 7 years ago leaving us with 99% of human DNA sequence about which our knowledge is still sparse

DNA stores an enormous amount of information used in the development and functioning of an entire organism. It works as a database of schematics needed to create proteins and ways of regulating their concentration in living cells. The size of such database ranges from hundreds of thousands nucleotides in bacteria to even hundreds of billions in vertebrates [1] making the analysis process a incredibly difficult and complex task.

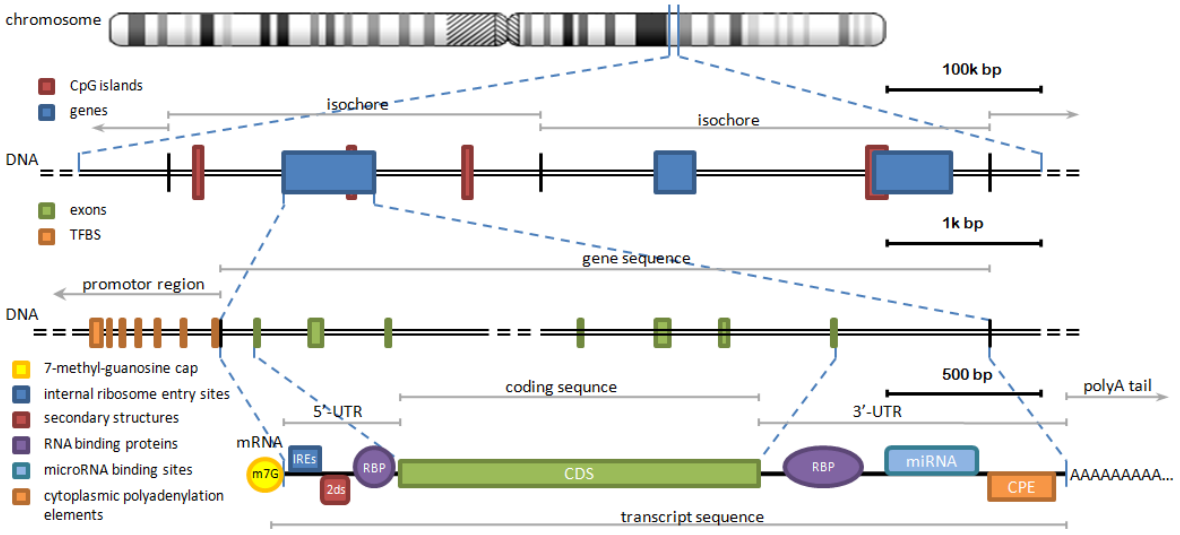


Fig.1. Functional elements of the DNA the lengths of specific sequence parts are based on average statistics made for human genome

Human DNA is organized into 23 chromosomes which consist of sequence fragments called isochores of over 300 kbp (kilo base pairs) long that significantly differ in G and C nucleotide content (GC%). CpG islands are much smaller genomic regions, on average 760 bp that contain a high frequency of CG dinucleotides believed to have a significant role in regulation of gene expression. Genes are DNA fragments (on average 56k bp) that are used in a process called transcription to form mRNA particle (transcript). Transcripts are much shorter than their respective genes (2,6kbp on average) since in a process called splicing large parts of the sequence (over 5kbp on average) called introns which separate exons (fragments of sequence coding proteins) are removed leaving only two uncoding fragments – 5'-UTR and 3'-UTR (on average 0,2kbp and 1kbp long) on both sides of the entire sequence coding protein structure - CDS.

Thanks to the huge amounts of currently available genomic data, available in public repositories the important role of structural and compositional features of various nucleotide sequence regions was recognized to play a very important role in gene regulation and expression – Figure 1.

One of the key players is the 3'-UTR containing many different regulatory elements like miRNA or protein binding domains which are responsible for mRNA turnover. In most of the vertebrates 3'-UTRs are significantly longer than their 5' counterparts, indicating their significant potential for regulation. In addition, the average length of 3'-UTR sequences has increased during the evolution, suggesting that their length might be related to organism complexity [2].

Another important fragment is the promoter region located upstream from the transcription initiation site which contains binding sites for proteins called transcription factors. These proteins are critical for making sure that genes are expressed in the right place, at the right time and in the right amount depending on the changing requirements of the organism. [3, 4]

Regulatory elements involved in many biological processes can be found in various genome parts - across the entire gene sequence and it's neighborhood. Their analysis involves hundreds of different applications and statistical methods for the identification of significant signal changes induced by their presence. The amount of different methods and applications used for DNA and RNA sequence analysis grows very rapidly. Most commonly used applications focus on the sequence similarity search between selected motifs and the target sequence leading to discoveries of transposons [5], promoter regions [6], miRNAs [7] and protein binding domains [8] or even gene similarities across various species [9]. The complexity level of the methods ranges from very low which check for the occurrence of specific sequence fragments to very high looking for closely related motifs, which in many cases share the same role, using position-weight-matrix models [10] or sequence alignment [11].

2. Aim of the research

The goal of the introduced experiment is to test the following hypothesis: "There is a correlation between the structural and functional features of the genes".

Tab.1.

Nucleotide sequence features tested in the experiment, numbers represent the amount of different values of selected feature tested for each sequence

Gene structure			Genome structure		
Feature name	values	description	Feature name	values	description
Sequence length	4	separately for: gene, 5'UTR, CDS, 3'UTR	Isochores	1	GC% of isochore with each gene
GC percentage	7	separately for: gene, 5'UTR, CDS, 3'UTR, prom1k/2k/5k upstream	Transposons	5x7	5 different families, separately for: Gene, 5'UTR, CDS, 3'UTR, prom1k/2k/5k upstream
Number of introns	1	for each gene sequence	Simple repeats	20x7	20 types, , separately for: gene, 5'UTR, CDS, 3'UTR, prom1k/2k/5k upstream
Intron/exon length	1	for each gene sequence	CpG islands	2	Separately for gene and prom 5k upstream
Polyadenylation sites	1	for each gene sequence	SNP (Single Nucleotide Polymorphism)	7	separately for: gene, 5'UTR, CDS, 3'UTR, prom1k/2k/5k upstream
Regulatory elements			Gene function		
TFBS (Transcription Factor Binding Sites)	75x7	75 different types, separately for: Gene, 5'UTR, CDS, 3'UTR, prom1k/2k/5k upstream	KEGG pathway	371	371 different functional pathways
RBP (RNA Binding Proteins) interaction sites	58x4	58 different types, separately for: gene, 5'UTR, CDS, 3'UTR	custom microarray expression profile study	6	changes of gene expression level 0,12 and 24 hours after irradiation of Me45 and K562 cells
microRNA binding sites	1101x3	1100 different types + sum of all, separately for: 5'UTR, CDS, 3'UTR	ArrayExpress: gene expression profile study	3651	3700 different threading factors

The hypothesis is very unspecific since it doesn't point out which out of known features should be included in the analysis making a perfect opportunity to test as much various factors as possible with hope of revealing unknown correlations.

Gene features can be derived by the use of our custom made software NucleoSeq and many other which are publicly available like miRanda [7] IsoFinder [12] or PatSearch [13] based on human genome data available on the University of California, Santa Cruz website: genome.ucsc.edu

Features assigned to individual transcripts, that will be taken into consideration can be divided into 4 separate classes according to Table 1. Most of the features are tested independently for various sequence regions including: entire gene sequence, 5'-UTR, CDS, 3'-UTR and promoter region 1000/2000/5000 bp upstream from the transcription initiation site.

When it comes to gene expression data the currently biggest database is the ArrayExpress, with 13685 submitted microarray experiments in nearly 7,3 TB of data as of May 2010. In this work only human genes are taken into account only the information about significant changes of their expression level due to 3651 submitted testing factors limiting the amount of data to 1,2GB.

3. Analysis workflow

Step 1: Gathering of gene sequence feature data based on public databases and available sequence analysis applications.

Sequence features described in Tab.1 are used to construct a table where rows represent individual Reference Sequence transcripts and columns their specific features – Tab.2. This is the most time consuming part of the experiment since there are many problems concerning sequence analysis. First of all there are no standard analysis workflows and the amount of available methods each with multiple sensitivity and specificity parameters is enormous. This requires many study hours in order to chose the most appropriate methods according to past researches described in the literature, and sometimes the need of testing the same features using different approaches choosing the best one afterwards.

Tab.2.

Part of the transcript feature table

RefSeq transcript ID	Gene: exon count	Gene: GC	Gene: ARE	Gene: CpG count	Isochore: GC	3'-UTR: miRNA binding sites	Me45 0h/C	K562 0h/C	ArrayExpress: rnai: p53 knockdown
NM_000014	36	37.7	0	0	38.95	6	-1	-1	0.0
NM_000015	2	36.1	6	0	39.19	11	0	-1	-1.0
NM_000017	10	52.7	0	2	49.27	4	-1	0	-1.0
NM_000018	20	55.1	0	1	54.41	18	-1	0	-1.0
NM_000019	12	51.9	3	1	40.10	27	0	0	0.0
NM_000021	12	50.2	0	1	42.25	57	1	0	-1.0
NM_000022	12	58.5	0	1	53.63	11	0	0	-1.0
NM_000023	10	57.7	0	0	55.92	14	0	0	-1.0
NM_000025	2	53.1	4	1	47.57	29	0	0	-1.0

When working with various data types originating from different sources there is a big risk that most of them might be unsynchronized with each other due to different time of preparation. For example information about location of regulatory sites downloaded from one website might be based on older sequences than those used in other parts of the research due to fact that they are updated every day. To prevent this all parts of the analysis should be performed based on the data from the same or very close release, which is sometimes impossible. Because of that one has to be aware the fact that a small fraction of the gene features might be unsynchronized with each other and that the results might become very quickly outdated.

What also causes many problems is the fact that some genes might have multiple values assigned for a single feature like when the same gene is located in more than one part of the genome or when there are multiple alternative transcript structures available, called the splicing variants. Therefore a god choose is to work on individual transcripts which are products of respective genes. This can also be problematic since in many cases the differences between alternative splicing variants only concern some small parts of them therefore many of the feature values become duplicated. Additionally many of the features cannot be analyzed for all genes of interest like those involving location of coding sequence, which is unavailable for many genes due to the fact that some play a role different from coding a protein structure.

The best choice would seem to gather as many features as possible to increase the probability of finding significant correlations. In this case the amount of data becomes a problem and not only a possibility since the size of the table containing all features of interest reaching over 8,200 rows and 46,000 columns.

To limit the amount of data and to unify the testing data size on the next step of the research, rows representing genes, that did not contain values for all chosen features were removed and features that will be impossible to interpret without additional data, like expression profile of individual patients.

Step 2: Testing statistical relationships between available gene features

To further limit the amount of data, and to extract only the information about relationships between the tested features, for each unique par a correlation measure has to be calculated. Due to the variability of tested features appropriate statistical methods need to be chosen when analyzing correlations. The feature values change in different intervals they are represented with different units and most importantly they form various distributions. To bypass this problem all features were binarized changing their value to 1 or 0 based on the following criteria:

- specific motif or feature exists in the analyzed sequence (when they are very rare)
- given statistic is higher or lower than mean value, like the GC percentage, sequence length or motifs occurring very frequently
- gene takes a part in the described process or is over/under expressed

The binarized feature values were then compared using the Chi-square test, but since Chi-square says only that there is a significant relationship between variables, and it does not say just how significant and important this is, additionally the Phi correlation measure was calculated. It's value ranges in an interval of -1 to 1 where 1 is the strongest positive correlation -1 negative and 0 means that the data are not correlated with each other.

When testing so many hypothesis in one experiment there is a big risk of committing the type I error depending on the chosen p-value cutoff level, therefore appropriate corrections for multiple testing must be applied. For this purpose the calculation of q-value used for determination of statistical significance instead of p-value was chosen according to [14].

Step 3: Removing insignificant and obvious correlations

In the next step only the significant correlations are taken into account based on calculated q-value and 0.05 significance level cutoff. The results of correlation study should be analyzed very carefully since some of them might turn out to be very obvious or well known, for example we might expect a very strong correlations between the occurrence of GC rich sequence motifs or CpG islands and GC rich parts of the sequence.

More complicated relations exist between the amount of transposons and the analyzed sequence length since because transposons are sometimes very long themselves their occurrence significantly increases the transcript length. Additionally long sequences tend to have much more functional elements which result from increased probability of occurrence by chance therefore the quantities of motifs should be presented in units independent from the sequence length, like occurrence per 1000 bases.

Some correlations might also occur due to specific experiment assumptions which is the most dangerous case. One of the examples is the correlation of CpG island amount between promoter region and the gene itself. Due to fact that CpG islands can be very long and since they were counted when they at least overlapped the analyzed region, in many cases the same CpG island could overlap both the promoter and gene region resulting in a high correlation score between those two sequence fragments. For this reason expected and insignificant correlations were removed reducing the amount of data to about 1000 rows representing only the correlations of two specific features. Some examples of them are gathered in Tab.3.

Tab.3.

Chosen correlations of transcript sequence features

Feature 1	Feature 2	Phi correlation	q-value
Gene: GC%	Isochore: GC%	0,714	<10^-9
Gene: GC%	CDS: GC%	0,773	
Gene: GC%	3'-UTR: miRNA binding sites	-0,448	
3'-UTR: ARE count	3'-UTR: miRNA binding sites	0,526	
CDS: Length	Gene: exon count	0,591	
Promotor: RELA TFBS	Promotor: NF-kappaB TFBS	0,486	
Irradiated Me45 0h	3'-UTR: miRNA binding sites	0,446	
Irradiated Me45 0h	Gene: GC	-0,645	
Irradiated Me45 0h	CDS: GC	-0,598	
Irradiated Me45 0h	3'-UTR: GC	-0,609	
Irradiated Me45 0h	Gene: iso_gc	-0,562	
Irradiated K562 0h	Promotor: CpG	0,391	
diseasestate: breast carcinoma	diseasestate: muscle invasive carcinoma	0,741	
diseasestate: brest carcinoma	Gene: GC%	-0,435	

Step 4: Choosing interesting and unknown correlations for further analysis

This is the hardest part of the research that requires human interpretation and cannot be easily done by a computer. Out of all correlations extracted in step 3 one has to chose the most interesting ones for detailed study.

As shown in table 3 very strong correlation was observed between gene and isochore GC percentage which confirms results presented in [15]. High overall gene GC percentage also results from the nucleotide composition of coding sequence (CDS) which due to specific codon bias can change in a large interval keeping the same protein structure information due to degeneracy of the genetic code, as was observed in [16]. Also a very strong positive correlation was observed between the gene GC percentage and amount of miRNA binding sites which also turns out to be already a known fact [17]

Long coding sequence is connected with large amount of exons which sounds obvious since they form in a process called splicing the final coding sequence. High occurrence of RELA transcription factor binding sites is connected with NF-kappaB sites but this only results from the fact that they are very similar and since the analysis algorithm specificity level allows slight differences between target sites and the original motif, the correlation might be expected.

What is very interesting is the negative correlations between the nucleotide content and changes of expression in irradiated Me45 cells. The largest correlation is observed for the entire gene GC% and since the GC% of gene is correlated with isochore GC% and negatively with miRNA binding sites amount such correlations are also observed with genes which expression changed due to irradiation. A slight decrease of correlation level was observed over the next 12h and 24h of experiment which is also similar to K562 cells expression profile. Additionally in K562 cells 1007 genes which expression increased due to irradiation out of 1208 have a CpG island in their promoter region which was not expected because of their negative correlation with GC rich genome region. This is explained by the correlation between promoter CpG islands and GC% of promoter and 5'-UTR sequence which is very high, while low with entire gene and isochore GC%.

Based on ArrayExpress expression profile data most of the genes expressed in breast cancer tissues are GC poor which is shown by the negative correlation measure. Additionally breast cancer expression profile shows huge similarities to muscle invasive cancer sharing the levels of expression profile changes for 1166 out of 1339 genes differently expressed due to at least one disease state.

Does it mean that the two cancers are very similar? Not necessarily, since even a few genes that are regulated in a different way can make a huge difference.

Step 5: Detailed analysis of selected factors

Features derived by the correlation study for irradiated Me45 cells were compared using all genes and not only those sharing all the analyzed features as described in step 1 and not based on binarized values. Statistical significance of differences in GC content and miRNA binding sites between up and down regulated genes due to irradiation was assessed with the Wilcoxon rank sum test for equal medians – Fig.2

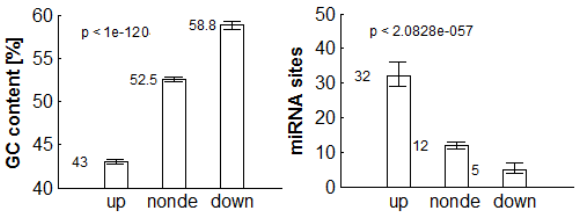


Fig.2. Median of GC% and miRNA binding sites in genes significantly up/down regulated or non differentially expressed due to irradiation in Me45 cells, p-value represents statistical significance between up and down regulated groups of genes

Step 6: Validation of results with different methods

What is pointed very often in many articles is that the statistical significance doesn't always mean biological significance because in many cases the analysis is possible to be performed only due to many simplifying assumptions. Additionally many of the analysis methods might not reflect the real biological nature of the processes therefore experimental verification is a must when one aims to publish the work in a journal with a high impact factor. Additional biological methods to confirm the identified relations are needed, like the real-time polymerase chain reaction used as a confirmation of imprecise microarray expression study. Further research including expression profiling of other cell lines and in other conditions also might confirm the observed relations although they overpass the borders of this work.

4. Conclusions

This work shows that lack of computational resources which is the main reason for rapid advancement in informatics involving parallel or heterogeneous computing isn't always the key problem when performing large scale analysis. The huge amount of different analysis scenarios and possibility to test hundreds of them even on a home computer can sometimes generate more results then

we are able to view and understand without the use of further analysis steps.

The amount of different factors and relations between them makes the genomic data mining and a very challenging task driven by the vision of huge benefits in many fields of science. Success will depend on our ability to properly interpret the data and overcome many problems related to data processing, which in turn requires us to adopt advances in bioinformatic methods.

Even if statistical significance does not necessarily mean biological significance, such large scale analysis may provide useful indication for further experimental work which normally could not be performed in such extent because of enormous coasts and lack of fast and efficient methods.

5. Bibliography And Authors

Bibliography

[1] Gregory, T.R., et al., *Eukaryotic genome size databases.* Nucleic Acids Res, 2007. 35(Database issue): p. D332-8.

[2] Mazumder, B., V. Seshadri, and P.L. Fox, *Translational control by the 3'-UTR: the ends specify the means.* Trends Biochem Sci, 2003. 28(2): p. 91-8.

[3] Latchman, D.S., *Transcription factors: an overview.* Int J Biochem Cell Biol, 1997. 29(12): p. 1305-12.

[4] Karin, M., *Too many transcription factors: positive and negative interactions.* New Biol, 1990. 2(2): p. 126-31.

[5] Kohany, O., et al., *Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor.* BMC Bioinformatics, 2006. 7: p. 474.

[6] Sandelin, A., W. Wasserman, and B. Lenhard, *ConSite: web-based prediction of regulatory elements using cross-species comparison.* Nucleic Acids Res, 2004. 32(Web Server issue): p. W249-52.

[7] John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. 2(11): p. e363.

[8] McCarthy, J.E. and H. Kollmus, *Cytoplasmic mRNA-protein interactions in eukaryotic gene expression.* Trends Biochem Sci, 1995. 20(5): p. 191-7.

[9] Margulies, E.H., et al., *Identification and characterization of multi-species conserved sequences.* Genome Res, 2003. 13(12): p. 2507-18.

[10] Bucher, P., *Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.* J Mol Biol, 1990. 212(4): p. 563-78.

[11] Larkin, M.A., et al., *Clustal W and Clustal X version 2.0.* Bioinformatics, 2007. 23(21): p. 2947-8.

[12] Oliver, J.L., et al., *IsoFinder: computational prediction of isochores in genome sequences.* Nucleic Acids Res, 2004. 32(Web Server issue): p. W287-92.

[13] Grillo, G., et al., *PatSearch: A program for the detection of patterns and structural motifs in nucleotide sequences.* Nucleic Acids Res, 2003. 31(13): p. 3608-12.

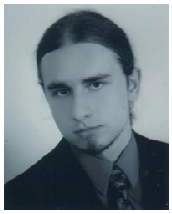
[14] Storey, J.D., *The positive false discovery rate: a Bayesian interpretation and the q-value.* Ann. Statist., 2003. 31(6): p. 2013-2035.

[15] Bernardi, G., *The human genome: organization and evolutionary history.* Annu Rev Genet, 1995. 29: p. 445-76.

[16] Costantini, M. and G. Bernardi, *The short-sequence designs of isochores from the human genome.* Proc Natl Acad Sci U S A, 2008. 105(37): p. 13971-6.

[17] Robins, H. and W.H. Press, *Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs.* Proc Natl Acad Sci U S A, 2005. 102(43): p. 15557-62.

Author:



MSc. Roman Jaksik
Institute of Automation
Silesian University of Technology
ul. Akademicka 16
44-100 Gliwice
tel. 502-102-169

email: *roman.jaksik@polsl.pl*

Acknowledgments:

The author would like to thank Joanna Rzeszowska-Wolny and Joanna Polańska for their many helpful comments and discussions regarding this work.

The work was supported by BW-445/RAu1/2010,t.7.