Tomas REKAŠIUS

AN EVOLUTIONARY MODEL FOR NONINFORMATIVE GENETIC SEQUENCES

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

1354

Vilnius VGTU LEIDYKLA TECHNIKA **2007**

VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Tomas REKAŠIUS

AN EVOLUTIONARY MODEL FOR NONINFORMATIVE GENETIC SEQUENCES

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

Vilnius LEIDYKLA TECHNIKA 2007

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2002–2006.

Scientific Supervisor

Assoc Prof Dr Marijus RADAVIČIUS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P)

The dissertation is being defended at the Council of Scientific Field of Mathematics at Vilnius Gediminas Technical University:

Chairman

Prof Dr Habil Leonas SAULIS (Vilnius Gediminas Technical University, Physical Sciences, Mathematics – 01P)

Members:

Prof Dr Habil Mindaugas BLOZNELIS (Vilnius University, Physical Sciences, Mathematics – 01P)

Prof Dr Habil Feliksas IVANAUSKAS (Vilnius University, Physical Sciences, Mathematics – 01P)

Prof Dr Habil Kęstutis KUBILIUS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P)

Prof Dr Habil Juozas KULYS (Vilnius Gediminas Technical University, Physical Sciences, Chemistry – 03P)

Opponents:

Prof Dr Kęstutis DUČINSKAS (Klaipėda University, Physical Sciences, Mathematics – 01P)

Prof Dr Habil Rimantas RUDZKIS (Institute of Mathematics and Informatics, Physical Sciences, Mathematics – 01P)

The dissertation will be defended at the public meeting of the Council of Scientific Field of Mathematics in the Senate Hall of Vilnius Gediminas Technical University at 11 a. m. on 2 March 2007.

Address: Saulėtekio al. 11, LT-10223 Vilnius, Lithuania

Tel.: +370 5 274 4952, +370 5 274 4956; fax +370 5 270 0112;

e-mail doktor@adm.vtu.lt

The summary of the doctoral dissertation was distributed on 2 February 2007.

A copy of the doctoral dissertation is available for review at the Library of Vilnius Gediminas Technical University (Saulėtekio al. 14, Vilnius, Lithuania) and the Library of the Institute of Mathematics and Informatics (Akademijos g. 4, Vilnius, Lithuania).

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Tomas REKAŠIUS

EVOLIUCINIS NEINFORMATYVIŲ GENETINIŲ SEKŲ MODELIS

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

Vilnius LEIDYKLA TECHNIKA 2007

Disertacija rengta 2002–2006 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas
doc. dr. Marijus RADAVIČIUS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).

Disertacija ginama Vilniaus Gedimino technikos universiteto Matematikos mokslo krypties taryboje:
Pirmininkas
prof. habil. dr. Leonas SAULIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, matematika – 01P).
Nariai:
prof. habil. dr. Mindaugas BLOZNELIS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Feliksas IVANAUSKAS (Vilniaus universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Kęstutis KUBILIUS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Juozas KULYS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, chemija – 03P).
Oponentai:
prof. dr. Kęstutis DUČINSKAS (Klaipėdos universitetas, fiziniai mokslai, matematika – 01P),
prof. habil. dr. Rimantas RUDZKIS (Matematikos ir informatikos institutas, fiziniai mokslai, matematika – 01P).

Disertacija bus ginama viešame Matematikos mokslo krypties tarybos posėdyje 2007 m. kovo 2 d. 11 val. Vilniaus Gedimino technikos universiteto senato posėdžių salėje.
Adresas: Saulėtekio al. 11, LT-10223 Vilnius, Lietuva.
Tel.: +370 5 274 4952, +370 5 274 4956; faksas +370 5 270 0112;
el. paštas doktor@adm.vtu.lt

Disertacijos santrauka išsiuntinėta 2007 m. vasario 2 d.
Disertaciją galima peržiūrėti Vilniaus Gedimino technikos universiteto bibliotekoje (Saulėtekio al. 14, Vilnius, Lietuva) ir Matematikos ir informatikos instituto bibliotekoje (Akademijos g. 4, Vilnius, Lietuva).
VGTU leidyklos „Technika“ 1354 mokslo literatūros knyga.

1. General Characteristic of the Dissertation

Topicality of the problem. During the past decades, due to the fundamental discoveries in the field of molecular biology, it has become a central subject of biology sciences. Previous focus of attention has been shifted from the identification of one specific gene to greater opportunities that have become possible by the sequencing of complete genomes. That, in turn, opened the door to the technologies of the so-called post-genomic era. They are often based on the computer analysis of the entire genome, i.e. on bioinformatics.

The numbers of nucleotides and amino acids in such databases of nucleotide sequences as *GenBank*, *DDBJ* or *EMBL* have been continuously increasing and have become enormous. With such extensive and continuously supplemented data amounts available, recognition of biological signals in an individual nucleotide sequences or the whole DNA, as well as their determination and visualisation of their function have become a complicated task and a relevant problem of bioinformatics. Almost all databases of bioinformatics have measures for visualisation of nucleo or amino acid sequences; besides, detailed "maps" of the entire genome are being concluded to facilitate researches. They allow having a detailed view of a small fragment of the sequence but you cannot see properties characteristic of only that sequence or visually distinguish them from the sequences that have other properties characteristic of them.

When the nucleotide or protein sequence of an unknown function is available, the usual practice in the determination of its purpose would be its comparison with the known sequences or protein structures. The methods used to determine a protein function or its structure in nucleotide sequences of a biological-genetic signal include the search for specific patterns to be compared with nucleotide or amino acids sequences (Waterman M. S. 1995). The next natural task is to determine functional or evolution relations among individual proteins, their groups and organisms, and to (re)construct phylogenetic trees (Gusfield D. 1997). However, such tasks need the measure of the distance between complex symbol sequences known.

Aim and tasks of the work. The research object is probabilistic properties of non-coding DNA (nucleotide) sequences. Available models of DNA sequences are reviewed and their basic assumptions are verified by statistical analysis of bacterial DNA sequences. On the ground of this analysis, the definition of non-informative genetic sequence is introduced and a mathematical model of “genetic noise” is proposed. Computer simulations of non-coding (non-informative) nucleotide sequence evolution are performed and

resulting sequences are compared with native ones. The task of visualisation of genetic sequences is an important part of the work. The main tasks of the work are the following:

1. to analyse the statistical features (independence, Markovity, long-range dependence, etc.) of bacterial DNA sequences, especially non-coding ones,
2. to formulate a definition of a non-informative nucleotide sequence ("genetic noise") and to propose its mathematical model,
3. using the methodology of functional data analysis and the distance metrics between oligonucleotides, to propose an efficient method for nucleotide sequence visualisation.

Scientific novelty and practical value. Until now, any randomised sequence of nucleotides or amino acids was considered to be a non-informative nucleotide or amino acid sequence. The work offers and substantiates the opinion that prior good knowing of biological-"genetic noise" is necessary to detect a biological signal in DNA sequences. There occurs a need for definition and accurate formulation of the notion of "genetic noise".

The statistical analysis carried out in the work reveals that the major part of even non-coding nucleotide sequences are not of the first order Markov chain, which is serious grounds for having doubts about the available models of nucleotide sequences, assumptions of their existence and adequacy of their application. This means that, for example, a comparison of real sequences with ones generated according to such models is not a reliable tool in the search either a biological signal or a biological function of a specific nucleotide (or amino acid) sequence. The same holds regarding the accuracy of phylogenetic trees reconstructed by means of these models. As an alternative for the existing models, a mathematical definition of non-informative nucleotide sequence or, in other words, of "genetic noise", has been formulated and its model has been proposed.

DNA of even very simple organisms – bacteria – is of a very long nucleotide sequence. Thus, its visualisation and presentation of achieved results is a topical issue. On the other hand, dealing with nucleotide sequences as sequences of categorical variables (e.g., by means of loglinear analysis) is complicated because of the large number of model parameters to be estimated.

In the work a new way to represent a nucleotide sequence as a real number is suggested. This representation should be "continuous" with respect to a "natural" distance between nucleotide sequences. For that, distances which take into account complexity of (binary) sequences are introduced. This representation obtained in this way offers an effective method for nucleotide sequence visualisation and analysis.

Methodology of research. The theory of discrete Markov fields is used to define a non-informative nucleotide sequence ("genetic noise") and to formulate its properties. The model of the non-informative sequence is verified by computer simulation of nucleotide sequence evolution. To analyse DNA sequence structure and nucleotide dependence, correlation and R/S (rescaled range) analysis are used (Beran J. 1994). To verify Markovity of a nucleotide sequence, loglinear and generalised *logit* models are applied and appropriate hypotheses are verified on their basis (Agresti A. 1990). For DNA visualisation methods of discrete mathematics and mutivariate analysis (principal components, factor analysis and multidimensional scaling) are used (Timm N. H. 2002). The data of bacteria genomes and the accompanying additional information has been taken from the *GenBank* database. In the course of the work, the methodology for the research of nucleotide sequence has been developed, a range of programmes necessary for statistical analysis and modelling of nucleotide sequences were written in the statistical analysis system *SAS*® environment.

Defended propositions

1. The models of genetic sequences usually used are based either on the assumption of independence of nucleotides or on the assumption of $k=1$ order Markovity. The investigation has revealed that this assumption is unfounded.
2. A simple evolution model of non-informative nucleotide sequence („genetic noise") has been proposed. According to the results, dependence of such sequences is of higher order k than 1, and, in general case, long-range dependence is their characteristic as well as for native nucleotide sequences.
3. In discrete mathematics, the distance between symbol sequences is usually defined as edit (Levenshtein) distance, which is not very suitable for distances between sequences with a complex structure of interaction of adjacent symbols. A different distance which can be treated as a discrete analogue of a well-known Sobolev norm and to a higher extent maintains information of the structure of DNA "words" under comparison has been introduced.
4. An efficient way for visualisation of nucleotide sequences has been proposed, which is free of disadvantages inherent to the traditional CGR (chaos game representation) "genome signature", for instance, its fractality. Pictures obtained are smooth and facilitate the comparison of all oligonucleotide combinations of length $n \leq 10$ in DNA sequences.

The scope of the scientific work. The scientific work starts with the general characteristic of the dissertation, introduction to statistical analysis of nucleotide sequences and review of literature. It consists of three chapters, conclusions, list of literature, list of publications and addenda. Dissertation is written in Lithuanian.

2. Contents

2.1. Introduction

The first chapter "Introduction to statistical analysis of nucleotide sequences" is aimed at presentation of DNA sequences, their features and features of individual nucleotides. DNA of any organism consists of two parts: gene coding and non-coding. DNA of a major part of both procariotic and eucariotic organisms is non-coding but its purpose is not fully clear. Evolutionary DNA models are usually developed to describe gene-coding (informative) sequences. However, from a mathematical point of view, in order to find a signal it is necessary to be well aware of what the noise is and what are its features. Because of the peculiarities of genome structure, non-coding bacterial DNA sequences are taken as an object of exploration of non-informative nucleotide sequences ("genetic noise").

Procariotic genomes are much more compact, they are free from repeated sequences, and almost all their genes are exclusively unique. The tendency of genes to make operons, a short distance between promotoric and regulation parts and between coding part reveal that by their meaning coding and non-coding DNA sequences differ a lot and inside they are more homogeneous than those in eucariotic genomes. An individual non-coding sequence of nucleotides is a handy object for the research on the most ordinary genome structure, the "grammar" of nucleotide sequence. An answer to the question what is the rule ("grammar") that generates a non-informative nucleotide sequence would facilitate finding out which DNA sequence is informative and how much it, as a biological signal, is important.

Further, this chapter gives a comprehensive description of evolutionary nucleotide sequence models. Traditionally such methods are divided into independent nucleotide models and context-dependent models. The first one is based on an assumption that nucleotides in the sequence evolve independently from each other (Hasegava, Kishino, Yano 1985). However, according to statistical analysis of real DNA sequences, this assumption is not substantiated. Recently, context-based mutation models have appeared (Arndt 2003, Hwang, Green 2004, Siepel, Haussler 2004). They are usually designed so that the stationary distribution of the evolution process of nucleotide sequences is

Markov; besides, the process is reversible in time (Jensen J. L. 2005). Thus, a short-range dependence should be a characteristic property of the DNA sequences. However, as it is shown in chapters 2 and 3, even non-coding sequences of bacteria possess long-range dependence. The chapter is ended by a synopsis of genetic databases, and raises the issue that is faced when looking for biologically important information in nucleotide sequences.

2.2. Model of Non-informative Nucleotide Sequence

Darwinian evolution is based on the principle of survival of the fittest. This is an optimization problem: one has to find an individual equipped with properties that are optimally suited to solve survival problems. This optimization becomes very hard in populations of limited size, but nature's strategy of optimizing life as we know it is extremely efficient and simple: increase variation on the basis of genotypes and select the phenotypes to decrease diversity. One of the ways to increase the variety of genotypes is mutations.

Let's define the fixed length sequence *x* of *n* symbols as follows:

$$x = x_1 x_2 ,..., x_n , \qquad x_l \in \mathrm{A}, l = \overline{1, n}, \tag{1}$$

where A is a finite set (alphabet). For DNA sequences $\mathrm{A} = \{A, C, G, T\}$. It is natural to consider, that DNA sequence evolution in time is described by a discrete time finite homogeneous Markov chain

$$X(t) = \{x_l(t), l = \overline{1, n}\}, \quad X(t) \in \mathrm{A}^n, T = \{0,1,2,...\}. \tag{2}$$

In this way, we have two evolution directions of the sequence *X*:

1) evolution in time

$$X(t) \longrightarrow X(t+1).$$

In the stationary case distribution of *X(t)* is independent of *t*. Thus, it defines probability distribution of a random sequence *X* on the set of sequences A^n and we can consider its

2) evolution in space

$$x_l \longrightarrow x_{l+1}.$$

In fact, it is not clear enough what a "non-informative nucleotide sequence" means. There is no "genetic noise", i.e. sequence, which definitely has no genetically important information which is necessary for survival of an organism. The definition of non-informative nucleotide sequence is based on the following assumptions.

1. Non-coding regions of DNA have not direct impact on survival of biological species and thus are not (so) genetically important,
2. Evolution of non-coding regions has simple structure and are controlled by local factors. For instance, in this work we ignore insertions and deletions and assume that probability of mutation in any site depends exclusively on its nearest neighbours.
3. Any part of genome (DNA sequence) can be significant for survival of species in non-stationary environment. Therefore only a stationary distribution of non-coding sequence evolution can be treated as "non-informative", i.e. as "genetic noise".

Definition

Let the evolution $X(t), t \in T$ *of nucleotide sequence* $X \in \mathrm{A}^n$ *in time be a (discrete time) homogeneous Markov chain with a given transition probabilities* Π *of a simple structure. If there exists its stationary distribution* Q *on* A^n*, a random sequence* X *with the distribution* Q *is called non-informative or "genetic noise".*

Assume for simplicity that the site state set $\mathrm{A} = \{0,1\}$ and consider the evolution in time of a random sequence $\{X(t), t \in T\}$, $X(t) \in \mathrm{A}^n$, of the length n. Suppose that this evolution is Markov and homogeneous in both time and space but in each site depends on its nearest neighbours. Namely,

$$\pi_{uzv} := P\{x_l(t+1) = \bar{z} \mid x_{[l-1,l+1]}(t) = uzv\}, \qquad (3)$$
$$l = \overline{2, n-1},\ u, z, v \in \mathrm{A},\ t \in T.$$

Here $\bar{z} = 1 - z$ and $x_{[l-1,l+1]} = x_{l-1}x_l x_{l+1}$ (we omit the argument t).

Let X denote the "noise" obtained by this evolution, i.e. X is a random sequence of 0's and 1's with the stationary (invariant) distribution of $\{X(t), t \in T\}$. It is completely determined by 8 scalar parameters $\pi := \{\pi_{uzv}\}$.

Proposition

If X is homogeneous Markov chain (in space) of order $k < n/4$ then:

a) $k = 1$, X is reversible and depends on 2 parameters, θ and γ,

b) the probabilities π satisfy equalities

$$\theta = \frac{\pi_{000}}{\pi_{010}}\ , \quad \gamma = \frac{\pi_{010}}{\pi_{000}} \frac{(\pi_{100} + \pi_{001})}{(\pi_{110} + \pi_{011})}, \quad \frac{\pi_{101}}{\pi_{111}} = \theta\gamma^2 .$$

Computer simulation. One of directions of DNA researches is an analysis of dependences in nucleotide sequences. Long-range dependence is characteristic even for non-coding regions of bacteria DNA sequences (Buldyrev at al. 1995). By means of computer simulations and R/S analysis it is revealed that sequences simulated according to the proposed evolution model demonstrate the similar long-range dependence behaviour as native ones.

Let $A = \{0,1\}$. The sequence (2) evolves under the context-dependent mutation model (Glauber dynamics) (3). Probability of nucleotide mutation in the sequence depends on two neighbouring nucleotides. Several different sets of transition probabilities are considered, for example: symmetric case where $\pi_{uzv} = \pi_{vzu}$, non-symmetric case where $\pi_{uzv} \neq \pi_{vzu}$. Simulation of the sequence evolution starts from a random binary sequence and 10^7 iterations (mutations) are performed. It is assumed that the sequence obtained has (approximately) stationary distribution.

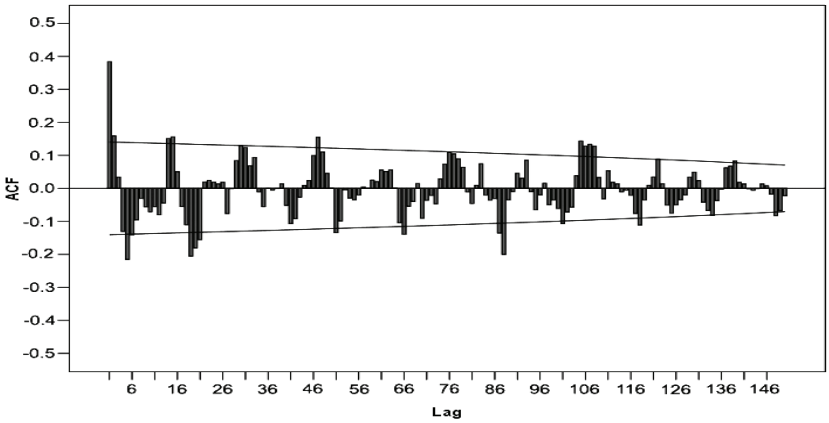

Fig 1. Autocorrelation function of simulated binary nucleotide sequence of length $n = 200$, non-symmetric transition probabilities

Let *X(t)* be a stationary process and there exists a real number $H \in (0.5,1)$ and a constant $c_p > 0$ such that autocorrelation function

$$\rho(k) \sim c_p \mid k \mid^{2H-2}, k \to \infty . \quad (4)$$

Then *X(t)* is called a stationary process with long-range dependence. The exponent *H* is called the Hurst parameter. To assess it, the R/S analysis is used.

For the sequence shown in figure 1, the estimated Hurst parameter value is $\hat{H} = 0.785$. For comparison, non-coding DNA sequences of bacteria *E.coli* are taken, the rule of recoding into the binary sequence is $\{C,G\} \to \{1\}$, $\{A,T\} \to \{0\}$. For the major part of bacteria the Hurst coefficient of non-coding sequences $\hat{H} \in (0.5,1)$. The same property with similar values of the Hurst coefficient holds also for the modelled sequences. Hence, it is shown that long-range dependence is a characteristic feature of non-informative nucleotide sequence, i.e. "genetic noise", generated by the simple evolution model with local interactions.

2.3. Statistical Analysis of Non-coding Bacterial DNA

This chapter is devoted to the statistical analysis of nucleotide sequences. The fact that nucleotides in the sequence are not independent is easily checked by chi-square independence test. However, what is the type of dependence? The aim of this part of the work is to verify the Markovity of bacterial DNA via loglinear analysis (Avery P. J. 2002). Generalized *logit* model is applied to verify the first order Markovity in all coding and non-coding, leading and lagging DNA strand sequences. The data we deal with is of the following form:

$$\{(y_l, z_l), l = \overline{1,N}\}, \quad (5)$$

where $y_l = x_{2l}$, $z_l = (x_{2l-1}, x2_{l+1})$, $l = \overline{1,N}$.

Assumptions:

1) $\{y_l, l = \overline{1,N}\}$ are conditionally independent given $\{z_l, l = \overline{1,N}\}$,
2) impact of *z's* on *y's* is homogeneous (does not depend on sites *l*).

This assumption is valid, in particular, if the sequence *X* is a homogeneous Markov chain. Thus, standard assumptions of regression models hold and we

can apply standard statistical software to perform statistical analysis. We use *SAS*® (proc **CATMOD**) to fit loglinear model to the data.

Let the state space be $\mathrm{A} = \{A, C, G, T\}$. A saturated *logit* model with the reference state 'T' is given by equality:

$$\log\left(\frac{P\{x_{[2l-1,2l+1]} = uzv\}}{P\{x_{[2l-1,2l+1]} = u'T'v\}}\right) = \lambda_0 + \lambda_{uz}{}^{L} + \lambda_{zv}{}^{R} + \lambda_{uzv}{}^{L\&R}, \qquad (6)$$

$u, z, v \in \mathrm{A},\ z \neq 'T'.$

For the Markov chain the interaction term $\lambda_{uzv}{}^{L\&R}$ should be zero, thus the task is to verify the hypotheses:

$$\begin{aligned} H_0 &: \quad \lambda_{uzv}{}^{L\&R} = 0 \qquad \forall u, z, v \in \mathrm{A} \\ H_1 &: \quad \lambda_{uzv}{}^{L\&R} \neq 0 \qquad otherwise. \end{aligned} \qquad (7)$$

For further analysis, DNA of bacteria *E.coli* was taken. The value of Likelihood Ratio (LR) statistic for full genome is LR = 18411.49, DF = 27, p-value < 0.00017.

Model (6) was fitted to each segment of the sequence separately. It turned out that the assumption of order *k = 1* Markovity is valid only for a small proportion of the segments (figure 2).

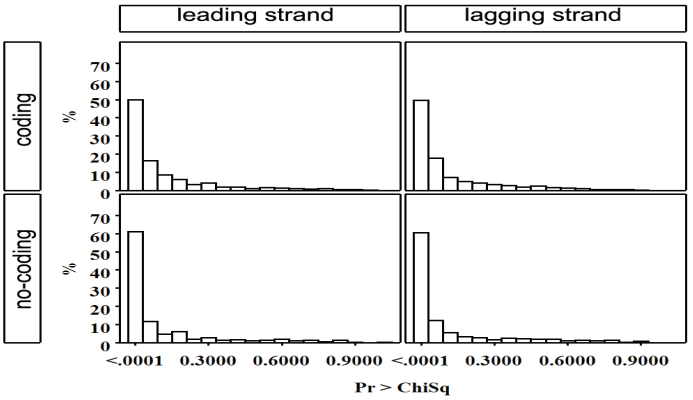

Fig 2. Distribution of p-values of LR statistic for testing the Markovity hypothesis

In general case, to describe the dependence among nucleotides in DNA chain the first order Markov model is insufficient. It is not difficult to write down a generalised theoretical *logit* function for a higher order Markov model, however the practical model assessment is complicated because of large number of its parameters.

2.4. Visualisation of nucleotide sequences

This chapter is devoted to visualisation of nucleotide sequences. DNA consists of four nucleotides briefly called *A*, *C*, *G*, *T*. Let us identify them with vertices of a square with coordinates (0,0), (1,1), (0,1), and (1,0), respectively. Thus, we have a natural isomorphism $\nu : \{A, C, G, T\} \to \mathrm{A} \times \mathrm{A}$, $\mathrm{A} = \{0,1\}$, where

$$\{\nu(A), \nu(C), \nu(G), \nu(T)\} = \{(0,0), (1,1), (0,1), (1,0)\}. \quad (8)$$

Chaos game representation (CGR) of DNA algorithm: a) recode DNA sequence into two binary sequences of the same length (8) and identify each of them with a fractional dyadic number, b) sequentially multiply these two numbers by 2 and take their fractional parts as a new pair of numbers, c) plot the pairs obtained on the graph. Resulting picture is called CGR "genome signature". "Genome signature" is an efficient way to picture long nucleotide sequences (Jeffrey H. J. 1990), but it has several undesirable features, and one of them is their fractality. For example the difference between dyadic representation in the interval [0,1) of two sequences *10000000000* and *01111111111* is less than 2^{-10} whereas the difference for "similar" sequences *00000000001* and *11111111110* is greater than $1\text{-}2^{-9}$. Due to the fractal character of DNA signatures the easily comprehensible Euclidean distance does not represent genetical similarity (dissimilarity) of oligonucleotides appropriately and hence the differences between them are difficult to interpret. To make "genome signature" to be appropriate for DNA visualisation the "natural" identification of nucleotide sequence with a (dyadic) point in the unit square should be replaced by a more subtle mapping. This mapping must be continuous in the sense that (genetically) "close" nucleotide sequences are represented by close points in the square and vice versa.

In bioinformatics, different genetic sequences are being compared rather frequently. Relationship of two organisms, when comparing their DNA, also gets to the calculation of the distance between two genomes. With such distances between the species known, phylogenetic trees could be (re)constructed, or the origin of species could be analysed. Further, the features

that should be characteristic of a measure of a distance between complicated sequences appropriate for such tasks are discussed.

Any finite sequence with elements from a finite alphabet can be identified with a rational number from the interval [0,1). In turn, any sequence of real numbers no matter how long it may be and any vector of a very large dimension can be treated as a function, i.e., merely as a point in a functional space. This is the paradigm of the functional data analysis (Silverman B.W. 1997). In this work we attempt to apply this approach to genetic (nucleotide) sequence visualisation and analysis. The work is continued by defining the distance between two binary sequences.

The distance proposed below is based on a operator of "differentiation". In some sense it is a discrete analog of Sobolev norm which is well known in the functional analysis. Assume for simplicity that the sequence *x* defined in (1) is binary, i.e. $A = \{0,1\}$. In the sequel *x* is identified with the corresponding vector in the space R^n. The difference ("differentiation") operator *B* is defined in the following way. Let:

$$M_n \xrightarrow{B_1^{(n)}} M_{n-1} \xrightarrow{B_1^{(n-1)}} M_{n-1} \to \ldots \to M_2 \xrightarrow{B_1^{2}} M_1, \qquad (9)$$
$$M_i \subset R^i,\ i = \overline{1,n}.$$

Here the operator $B_1^{(k)},\ k \in \{2,\ldots,n\}$ is expressed by the formula

$$B_1^{(k)} x = \{(x_{i+1} - x_i)/2, i = \overline{1,k-1}\}, \qquad (10)$$

and the operators $B_l^{(n)} : M_n \to M_{n-l}$ are obtained recurrently from the formula

$$B_l^{(n)} = B_1^{(n-l+1)} B_{l-1}^{(n)}, l = \overline{2,n-1}. \qquad (11)$$

The upper index *n* of the operator $B_l^{(n)}$ indicates the dimension of the space it acts in, while the lower index *l* shows the extent to which the dimension of its mapping is smaller. For a given $x \in M_n \subset R^n$, coordinates of the corresponding point $y = Bx$ in the space $R^{n(n+1)/2}$ are expressed by the formula

$$y = y(x) = Bx = (x, B_1^{(n)} x, B_2^{(n)} x, \ldots, B_{n-1}^{(n)} x). \qquad (12)$$

Let a positive defined diagonal matrix $W = \{w_1,...,w_q\}$ of the order $q=n(n+1/2)$ be given. Define the weighted inner product in R^q by the equality

$$(u,v)_W := u^T W v = \sum_{i=1}^{q} w_i u_i v_i \tag{13}$$

and set $|u|_W = \sqrt{(u,u)_W}$. The distance $d = d_W$ between two binary sequences x and z is defined as

$$d(x,z) = |y(x) - y(z)|_W,\ x,z \in M_n,\ y(x), y(z) \in M \subset R^q. \tag{14}$$

Define the cyclic shift operator $T: M_n \to M_n$ by the equality

$$T(x) = x_n x_1 x_2 ... x_{n-1},\ x \in M_n. \tag{15}$$

For a given positive sequence $\rho_0, \rho_2, ..., \rho_{n-1}$ define average (smoothed) distance

$$d_\rho(x,z) = \sum_{j=0}^{n-1} \rho_j (T^j(x), T^j(z)). \tag{16}$$

Given a matrix ***D*** consisting of the pairwise dissimilarities of the sequences x (in the space R^q) the goal is to reduce the dimensionality q of the data set to a sufficiently small value so that the distances between the sequences x in the low dimension space would be as close to the original distances as possible. It is a classical problem of multidimensional scaling. To solve it, here we apply *SAS*® procedure **MDS** (table 1-2, figure 3).

Table 1. One-dimensional projections $\varphi(x)$ of n = 5 long sequences x. Distance $d(x,z)$

No	x	$\varphi(x)$	No	x	$\varphi(x)$	No	x	$\varphi(x)$	No	x	$\varphi(x)$
1	10101	0.0000	9	10110	0.2588	17	00000	0.5083	25	10010	0.7755
2	00101	0.0403	10	11100	0.2969	18	01111	0.5523	26	11000	0.7799
3	10100	0.0623	11	10001	0.3307	19	11110	0.5743	27	00010	0.8224
4	00100	0.1033	12	00110	0.3550	20	11001	0.6193	28	01000	0.8481
5	11101	0.1519	13	01100	0.3807	21	10011	0.6450	29	11011	0.8967
6	10111	0.1776	14	00001	0.4257	22	01110	0.6693	30	01011	0.9377
7	01101	0.1988	15	10000	0.4477	23	01001	0.7155	31	11010	0.9597
8	00111	0.2458	16	11111	0.4917	24	00011	0.7288	32	01010	1.0000

Table 2. One-dimensional projections $\varphi(x)$ of $n = 5$ long sequences x. Distance $d_\rho(x,z)$

No	x	$\varphi(x)$	No	x	$\varphi(x)$	No	x	$\varphi(x)$	No	x	$\varphi(x)$
1	11011	0.0000	9	01101	0.2345	17	00110	0.5302	25	10010	0.7995
2	11101	0.0322	10	10110	0.2672	18	01100	0.5574	26	10100	0.8396
3	10111	0.0582	11	11010	0.2946	19	00011	0.5850	27	00001	0.8604
4	11111	0.1076	12	01110	0.3599	20	11000	0.6112	28	10000	0.8837
5	01111	0.1163	13	00111	0.3888	21	10001	0.6401	29	00000	0.8924
6	11110	0.1396	14	11100	0.4150	22	01001	0.6997	30	00010	0.9418
7	10101	0.1532	15	11001	0.4426	23	00101	0.7385	31	01000	0.9678
8	01011	0.1948	16	10011	0.4698	24	01010	0.7784	32	00100	1.0000

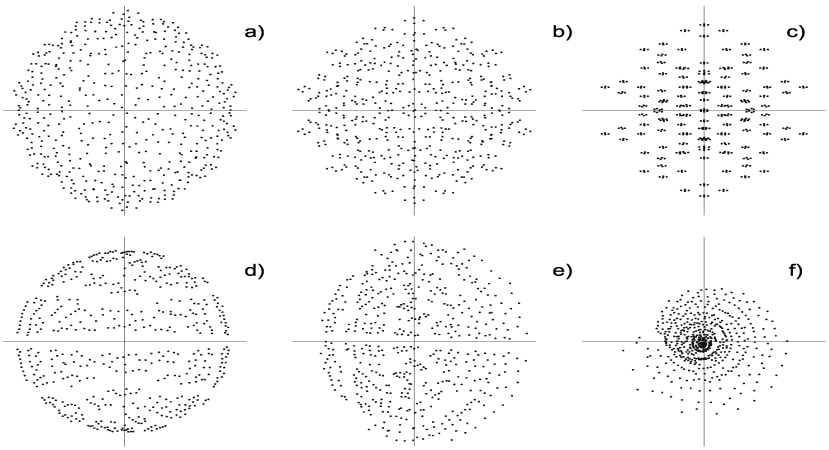

Fig 3. Two-dimensional projections of $n = 9$ long binary sequences x.
Received x projections depend on weight matrix W and distance $d(x,z)$ (a, b, c),
smoothed distance $d_\rho(x,z)$ (d, e, f)

Modified "genome signature". DNA sequence $S = \{s_l, l = \overline{1,m}\}$, $S \in \mathrm{A}^m$, can be expressed in an equivalent form as two-dimensional binary sequence (8). Let $\varphi(x)$ be an one-dimensional projection of x (table 1-2). Attributing the coordinate $\varphi(s(j))$ to the moving binary sequence $s(j) = s_j s_{j+1},...,s_{j+n}$ of the length n for the entire DNA we obtain a set of two-dimensional points. This set is called "genome signature".

Using bacterial DNA data from *Genbank*, we have discovered some characteristic patterns presented below (figure 4). The signature of the sequence of 10^5 nucleotides is drawn. Differently from traditional "genome signature" obtained by CGR, the patterns obtained are rather smooth, the set of points is not divided into sub-squares and fractality is not so evident.

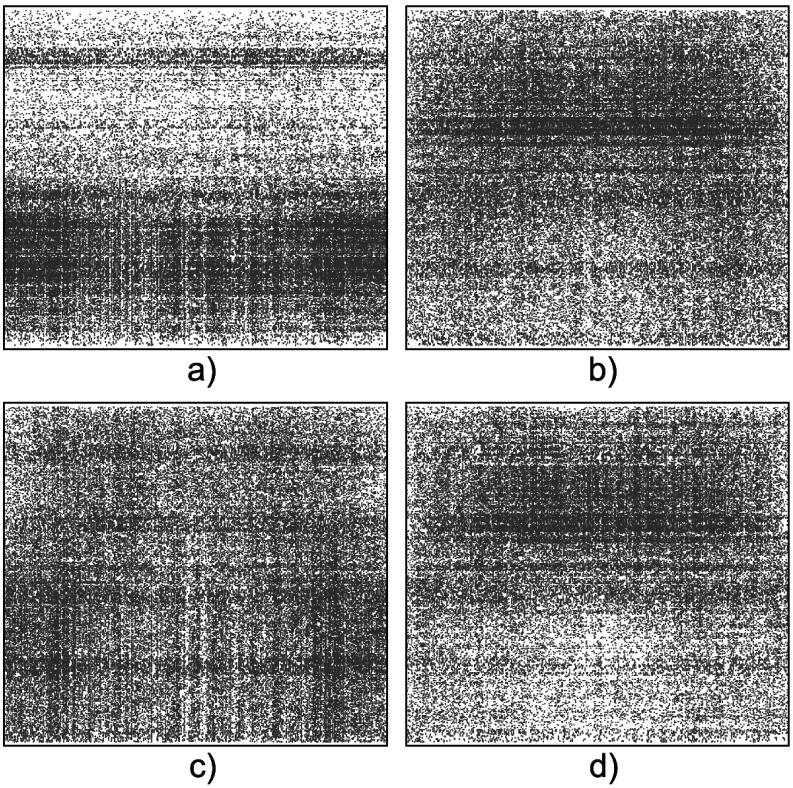

Fig 4. Modified bacterial "genome signature": *Bordetella bronchiseptica* (a), *Coxiella burnetii* (b), *Escherichia coli K12* (c), *Helicobacter pylori J99* (d)

As it could be expected, related bacteria gave similar genome signature. It is clear that such genome signature patterns depend also on one-dimensional code distribution in the sequence. Taking only one genome signature coordinate a DNA sequence can be analysed by one of two characteristics of nucleotides.

3. General Conclusions

1. The probability model of non-informative nucleotide sequence or, in other words, "genetic noise" (an analogue of the "white noise") is proposed and its properties are studied mainly by computer simulation. The long-range dependence in DNA sequences has been extensively studied and is considered as an evidence of their complexity and hierarchical structure. The work reveals that the long-range dependence is intrinsic property of a the "genetic noise" generated by a simple evolution model.
2. Common probabilistic models of genetic sequences usually assume either independency of nucleotides or the first order Markovity. The analysis of dependences of bacteria genomes in non-coding nucleotide sequences has been carried out and hypotheses of independence and Markovity in these sequences have been tested. It have been shown that these hypotheses are rejected in a major part of the sequences.
3. A new measure of the distance between nucleotide sequences and the efficient method for visualisation of nucleotide sequences have been proposed. It facilitates visualisation of a long DNA sequence via a compact and smooth picture of all oligonucleotides of length up to 10 in the sequence.

Published works on the topic of the dissertation
In the acknowledged editions

1. ŽIDANAVIČIŪTĖ, J.; REKAŠIUS, T. Fitting Markov property to genetic sequences. *Liet. matem. rink.*, **46**, spec. release 2006, p. 280–285 (in Lithuanian).
2. REKAŠIUS, T. Visualisation of nucleotide sequences. *Liet. matem. rink.*, **46**, spec. release 2006, p. 390–396 (in Lithuanian).
3. REKAŠIUS, T. Research on dependencies in simulated DNA sequences. *Liet. matem. rink.*, **45**, spec. release 2005, p. 363–368 (in Lithuanian).
4. REKAŠIUS, T.; TIMINSKAS, A. Statistical analysis on avoidance of palindromic sequences in genomes of microorganisms. *Liet. matem. rink.*, **43**, spec. release 2003, p. 649–654 (in Lithuanian).

Acknowledgement

I am very grateful to Professor A.Timinskas (Institute of Biotechnology) for introducing me to bioinformatics and stimulating discussions and Professor G. Kulldorff (University of Umeå, Sweden) for the opportunity to be on placement at the University of Umeå.

About the author

Tomas Rekašius was born in Rietavas, on 30 of September 1978.

First degree in Applied Mathematics, Faculty of Fundamental Science, Kaunas University of Technology, 2000. Master of Applied Statistics, Faculty of Fundamental Sciences, Vilnius Gediminas Technical University, 2002. In 2002–2006 – PhD student of Vilnius Gediminas Technical University, Department of Mathematical statistics. Tomas Rekašius in 2006 was on internship at Umea university, Department of Mathematical Statistics. 2006–2007 – Assistant in Mathematical statistics Department of Vilnius Gediminas Technical University.

EVOLIUCINIS NEINFORMATYVIŲ GENETINIŲ SEKŲ MODELIS

Mokslo problemos aktualumas

Turint didelius ir nuolat papildomus genetinių duomenų kiekius, biologinių signalų atskirai paimtoje nukleotidų sekoje ar visoje DNR atpažinimas, funkcijos nustatymas ir vizualizavimas tampa sunkiai aprėpiamu uždaviniu ir labai aktualia bioinformatikos problema.

Darbo tikslas ir uždaviniai

1. Ištirti bakterijų DNR koduojančių ir nekoduojančių sekų statistines savybes (nepriklausomumas, markoviškumas) ir gautų rezultatų pagrindu pasiūlyti neinformatyvios nukleotidų sekos modelį.
2. Suformuluoti neinformatyvios nukleotidų sekos („genetinio triukšmo“) apibrėžimą ir pasiūlyti jos matematinį modelį.
3. Remiantis funkcionalinės duomenų analizės metodologija ir atstumo tarp oligonukleotidų tarpusavio atstumo matais pasiūlyti efektyvų metodą nukleotidų sekų vizualizavimui.

Mokslinis naujumas

1. Sudarytas *neinformatyvios* nukleotidų sekos matematinis modelis.
2. Pasiūlytas naujas nukleotidų sekų vizualizavimo būdas.

Tyrimų metodika

Neinformatyvios nukleotidų sekos apibrėžimui suformuluoti naudojama diskrečių Markovo laukų teorija. Modelis tikrinamas kompiuteriu modeliuojant nukleotidų sekos evoliuciją. DNR sekų struktūrai ir nukleotidų priklausomybėms jose tirti taikomi įprastos ir ilgą atmintį turinčios procesų statistikos metodai: koreliacinė ir R/S analizė, požymių nepriklausomumo

testas. Nukleotidų sekos markoviškumo hipotezei patikrinti naudojami logtiesiniai ir apibendrinti *logit* modeliai. DNR vizualizavimo uždaviniui spręsti taikomi daugiamatės statistikos metodai: pagrindinių komponenčių ir daugiamačio mastelio parinkimo metodai, faktorinė analizė. Reikalingi duomenys, t. y. bakterijų genomai ir juos lydinti papildoma informacija gauta iš *GenBank* duomenų bazės. Darbo metu buvo kuriama sava nukleotidų sekų tyrimo metodika, programinio statistinių tyrimų paketo *SAS*® aplinkoje parašyta visa eilė nukleotidų sekų statistinei analizei ir modeliavimui skirtų programų.

Praktinė vertė

Gautas neinformatyvios nukleotidų sekos modelis gali būti pritaikomas informatyvios sekos kaip biologinio signalo DNR sekose paieškai. Pasiūlytas atstumo tarp binarinių sekų matas gali būti pritaikytas DNR vizualizavimui, o tuo pačiu ir organizmų identifikavimui.

Ginamieji teiginiai

1. Dažniausiai naudojami genetinių sekų modeliai remiasi nukleotidų nepriklausomumo ar $k = 1$ eilės markoviškumo prielaida. Tyrimais parodyta, jog ši prielaida nepagrįsta.
2. Pasiūlytas nesudėtingas evoliucinis neinformatyvios nukleotidų sekos modelis. Kaip rodo rezultatai, tokios sekos priklausomybė aukštesnės eilės nei $k = 1$, o bendru atveju jai, kaip ir natyvioms nukleotidų sekoms, būdinga ilga priklausomybė.
3. Diskrečioje matematikoje atstumas tarp simbolių sekų apibrėžiamas kaip „redagavimo“ atstumas, kuris nelabai tinka atstumams tarp seku su sudėtinga gretimų simbolių tarpusavio sąveikos struktūra. Pasiūlytas kitas atstumas, kurį galima traktuoti kaip diskretų gerai žinomos Sobolevo normos analogą ir kuris geriau išlaiko informaciją apie lyginamų žodžių struktūrą.
4. Pasiūlytas efektyvus nukleotidų sekų vizualizavimo būdas neturintis tradiciniam „genomo parašui“ būdingų trūkumų (pirmiausia fraktališkumo), pasižymi glodumu ir leidžiantis nesunkiai palyginti visas iki $n = 10$ ilgio oligonukleotidų kombinacijas DNR sekoje.

Darbo apimtis

Darbą sudaro bendra darbo charakteristika, įvadinis skyrius ir literatūros apžvalga. Toliau seka trys skyriai. Pirmame iš jų suformuluojamas neinformatyvios nukleotidų sekos apibrėžimas ir pateikiamas matematinis modelis. Sekančiame skyriuje tiriamas nukleotidų sekų markoviškumas.

Paskutinis skyrius skirtas nukleotidų sekų vizualizavimui. Darbas baigiamas išvadomis, literatūros ir publikacijų sąrašais bei priedais.

Darbo išvados

1. Pasiūlytas neinformatyvios nukleotidų sekos, arba kitaip – genetinio triukšmo („balto“ triukšmo analogo) tikimybinis modelis, atliktas jo tyrimas modeliavimo būdu. Paprastai laikoma, kad DNR sekos dėl joms būdingos ilgos priklausomybės yra sudėtingos, hierarchinę struktūrą turinčios sistemos. Darbe parodyta, kad paprasčiausio modelio generuotai neinformatyviai nukleotidų sekai taip pat būdinga ilga priklausomybė.
2. Atlikta bakterijų genomų priklausomybių nekoduojančiose nukleotidų sekose analizė, ištirta tokių sekų markoviškumo savybė. Dažniausiai naudojamų genetinių sekų modeliai remiasi nukleotidų nepriklausomumo ar $k = 1$ eilės markoviškumo prielaida. Atliktais tyrimais parodyta, jog ši prielaida nepagrįsta.
3. Pasiūlytas atstumo tarp nukleotidų sekų matas ir efektyvus nukleotidu sekos vizualizavimo būdas leidžiantis gauti kompaktiškus ilgų sekų atvaizdavimus, kuris gerai atspindi visas iki $n = 10$ ilgio oligonukleotidų kombinacijas DNR sekoje.

Padėka

Už nuolatinę pagalbą ir konsultacijas dėkoju dr. Albertui Timinskui (Biotechnologijos institutas) bei profesoriui Gunnar Kulldorff už galimybę stažuotis Umeå universitete.

Trumpos žinios apie autorių

Tomas Rekašius gimė 1978 m. rugsėjo 30 d. Rietave.

2000 m. įgijo taikomosios matematikos bakalauro laipsnį Kauno technologijos universiteto Fundamentaliųjų mokslų fakultete. 2002 m. įgijo taikomosios statistikos magistro laipsnį Vilniaus Gedimino technikos universiteto Fundamentinių mokslų fakultete. 2002–2006 m. – Vilniaus Gedimino technikos universiteto Matematinės statistikos katedros doktorantas. Tomas Rekašius 2006 m. stažavosi Švedijos Umeå universiteto Matematinės statistikos katedroje. Nuo 2002 m. metu dirba asistentu Vilniaus Gedimino technikos universiteto Matematinės statistikos katedroje.

Tomas Rekašius

AN EVOLUTIONARY MODEL FOR NONINFORMATIVE GENETIC SEQUENCES

Summary of Doctoral Dissertation
Physical Sciences, Mathematics (01P)

Tomas Rekašius

EVOLIUCINIS NEINFORMATYVIŲ GENETINIŲ SEKŲ MODELIS

Daktaro disertacijos santrauka
Fiziniai mokslai, matematika (01P)

2007 02 02. 1,5 sp. l. Tiražas 100 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“, Saulėtekio al. 11, LT-10223 Vilnius
Spausdino UAB „Biznio mašinų kompanija“,
J. Jasinskio g. 16A, LT-01112 Vilnius