P r o c e e d i n g s

of the

2nd International Conference "Theoretical Approaches to BioInformation Systems" (TABIS.2013)

September 17 – 22, 2013, Belgrade, Serbia

Editors

B. Dragovich, R. Panajotović, D. Timotijević

Institute of Physics

Belgrade, 2014, SERBIA

Autor: Grupa autora

Naslov: 2nd INTERNATIONAL CONFERENCE
"THEORETICAL APPROACHES TO BIOINFORMATION
SYSTEMS" (TABIS.2013)
(Druga medjunarodna konferencija "Teorijski pristupi bioinformacionim
sistemima" - TABIS.2013)

Izdavač: Institut za fiziku, Beograd, Srbija

Izdanje: Prvo izdanje

Štampar: Ton Plus, Beograd

Tiraž: 100

ISBN: 978-86-82441-40-3

1. Dragović Branko

CIP – Katalogizacija u publikaciji
Narodna biblioteka Srbije, Beograd

PREFACE

Despite many remarkable successes in modern biology, living systems are not well understood comparing them to the ordinary physical systems. Living matter is much more complex and its investigation needs multidisciplinary approaches – including physics, mathematics, chemistry, computer science and some other related fields of sciences, in addition to biology. Probably the most complex and significant systems of the whole universe are bioinformation ones. Their studies need even more wide approaches, including information theory, complexity, networks, ... It is evident that some new methods should be discovered in theoretical approaches to model bioinformation systems. For example, standard mathematical methods which are so effective in physics are as well ineffective in biology, hence some new tools have to be invented.

This book contains proceedings of the 2nd International Conference "Theoretical Approaches to BioInformation Systems" (TABIS.2013), which was held 17–22 September 2013 in the Institute of Physics, Belgrade, Serbia. There were about 60 participants from 10 countries, who are scientists from different fields of theoretical research – physics, mathematics, computer science, biology and chemistry. Providing stimulating scientific ambient, participants had opportunity for exchange of ideas and discussion of recent results in the genomics, proteomics, cognitive neuroscience, networks and related subjects. Conference topics incuded:

- Structure and function of DNA and RNA
- Structure, function and interaction of proteins
- Gene expression and the genetic code
- Life as information processing
- Bioinformatics
- Neuron structure and signal processing
- Data mining and machine learning
- Cognitive modeling
- Networks

and some related topics.

Participants of the conference TABIS.2013 expressed their satisfaction with organization of this kind of scientific conferences and proposed also organization of the corresponding scientific network. We hope these proceedings will be useful not only to participants of TABIS.2013 but also to all who are interested in new theoretical approaches to bioinformation systems. There is a plan to organize next conference TABIS.2016.

We wish to thank all authors of articles in these proceedings as well as all speakers and participants of TABIS.2013. We are grateful to the Ministry of Education, Science and Technological Developments of the Republic of Serbia for a financial support. More information on TABIS.2013 is available at its internet page http://www.tabis2013.ipb.ac.rs/ .

E d i t o r s

Branko Dragovich
Radmila Panajotović
Dejan Timotijević

2nd INTERNATIONAL CONFERENCE "THEORETICAL APPROACHES TO BIOINFORMATION SYSTEMS"

TABIS.2013

Belgrade, September 17 - 22, 2013

Organizers

- Institute of Physics, University of Belgrade, Serbia
- Faculty of Mathematics, University of Belgrade, Serbia
- Faculty of Biology, University of Belgrade, Serbia

International Advisory Committee

Pavle Andjus (Belgrade, Serbia)
Miloš Beljanski (Belgrade, Serbia)
Radu Constantinescu (Craiova, Romania)
Janos Kertezs (Budapest, Hungary)
Andrei Khrennikov (Vaxjo, Sweden)
Sergei Kozyrev (Moscow, Russia)
Fionn Murtagh (London, UK)
Zoran Obradović (Temple, USA)
Sergey Petoukhov (Moscow, Russia)
Nataša Pržulj (London, UK)
Zoran Rakić (Belgrade, Serbia)
Miloje Rakočević (Belgrade, Serbia)
Miljko Satarić (Belgrade, Serbia)
Brunello Tirozzi (Roma, Italy)
Peter Tompa (Brussels, Belgium)
Alessandro Treves (Trieste, Italy)
Edward Trifonov (Haifa, Israel)

Scientific Committee

Marko Djordjević (Belgrade, Serbia)
Branko Dragovich (Belgrade, Serbia)
Michele Caselle (Torino, Italy)
Radmila Panajotović (Belgrade, Serbia)
Gordana Pavlović - Lažetić (Belgrade, Serbia)
Paul Sorba (Annecy, France)
Nenad Švrakić (Belgrade, Serbia)
Bosiljka Tadić (Ljubljana, Slovenia)
Igor Volovich (Moscow, Russia)

Local Organizing Committee

Branko Dragovich, Chairman (Institute of Physics, Belgrade)
Sanja Ćirković (Institute of Physics, Belgrade)
Mihajlo Mudrinić (Institute of Physics, Belgrade)
Radmila Panajotović (Institute of Physics, Belgrade)
Dragan Savić (Institute of Physics, Belgrade)
Nenad Švrakić (Institute of Physics, Belgrade)
Dejan Timotijević (Institute of Physics, Belgrade)
Radomir Žikić (Institute of Physics, Belgrade)

Sponsor

Ministry of Education, Science and Technological Development of the Republic of Serbia

Conference Participants

Anastasia Anashkina, Moscow, Russia
Pavle Andjus, Belgrade, Serbia
Ivana Antonijević, Belgrade, Serbia
Miloš Beljanski, Belgrade, Serbia
Svetlana Bojić, Belgrade, Serbia
Vladimir Čadež, Belgrade, Serbia
Michele Caselle, Torino, Italy
Viktor Cerovski, Belgrade, Serbia
Radu Constantinescu, Craiova, Romania
Ivan Dimitrijević, Belgrade, Serbia
Magdalena Djordjević, Belgrade, Serbia
Marko Djordjević, Belgrade, Serbia
Suzana Djurdjević, Belgrade, Serbia
Nikola Dragić, Belgrade, Serbia
Branko Dragovich, Belgrade, Serbia
Dragana Dudić, Belgrade, Serbia
Djordje Francuski, Belgrade, Serbia
Branislava Gemović, Belgrade, Serbia
Momir Glušica, Belgrade, Serbia
Tasko Grozdanov, Belgrade, Serbia
Jelena Grujić, Belgrade, Serbia
Jelena Guzina, Belgrade, Serbia
Miloš Jovanović, Belgrade, Serbia
Lajos Kalmar, Budapest, Hungary
Jovana Kovačević, Belgrade, Serbia
Claire Lesieur, St Julien en Genevoix, France
Dušan Malenov, Belgrade, Serbia
Vladimir Marković, Belgrade, Serbia
Ana Mijalković, Belgrade, Serbia
Miloš Milovanović, Belgrade, Serbia
Nenad Mitić, Belgrade, Serbia
Nataša Mišić, Belgrade, Serbia

Mihajlo Mudrinić, Belgrade, Serbia
Zoran Obradović, Temple, USA
Sanja Pajić, Belgrade, Serbia
Tanja Pajić, Belgrade, Serbia
Vesna Pajić, Belgrade, Serbia
Radmila Panajotović, Belgrade, Serbia
Gordana Pavlović-Lažetić, Belgrade, Serbia
Vladimir Perović, Belgrade, Serbia
Sergey Petoukhov, Moscow, Russia
Stevan Prodanović, Belgrade, Serbia
Nataša Pržulj, London, UK
Draginja Radošević, Belgrade, Serbia
Zoran Rakić, Belgrade, Serbia
Miloje Rakočević, Belgrade, Serbia
Igor Salom, Belgrade, Serbia
Paul Sorba, Annecy, France
Suzana Stanisavljević, Belgrade, Serbia
Ana Stanojević, Belgrade, Serbia
Neven Šumonja, Belgrade, Serbia
Saša Sviković, Belgrade, Serbia
Nenad Švrakić, Belgrade, Serbia
Jelena Tabas, Belgrade, Serbia
Bosa Tadić, Ljubljana, Slovenia
Dejan Timotijević, Belgrade, Serbia
Alessandro Treves, Trieste, Italy
Edward Trifonov, Haifa, Israel
Dušan Veljković, Belgrade, Serbia
Milan Vukićević, Belgrade, Serbia
Aleksandra Vukojičić, Belgrade, Serbia
Slobodan Zdravković, Belgrade, Serbia
Radomir Žikić, Belgrade, Serbia

Contents

A New Paradigm of Protein Structural Organization

Alexei N. Nekrasov [a]
Shemyakin-Ovchinnikov Institute of Bioorganic chemistry Russian Academy of Sciences

Anastasia A. Anashkina [b]
Engelhardt Institute of Molecular Biology Russian Academy of Sciences

Alexei A. Zinchenko [c]
Shemyakin-Ovchinnikov Institute of Bioorganic chemistry Russian Academy of Sciences

Abstract

The paper describes a new method for the analysis of protein sequences - the method of analysis of the information structure (ANIS method). The method uses a new approach to describe amino acid sequences and identify hierarchically organized elements in the information structure of protein sequences. It was shown that the top-level information structure elements correspond to topologically stable elements of the three-dimensional structure (structural domains). A new approach for the identification of functionally important protein fragments was proposed

[a] e-mail address: alexei_nekrasov@mail.ru,117997, GSP-7, ul. Miklukho-Maklaya, 16/10, Moscow, Russia

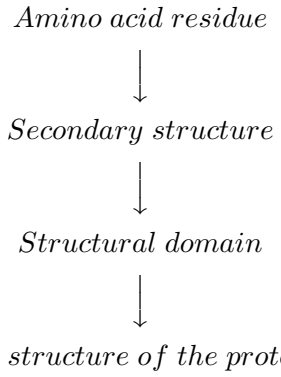
[b] e-mail address: nastya@eimb.ru,119991, GSP-1, ul. Vavilova, 32, Moscow, Russia

[c] e-mail address: alezina@mail.ru, 117997, GSP-7, ul. Miklukho-Maklaya, 16/10, Moscow, Russia

based on the ANIS method. The approach was tested in the protein engineering experimental studies. Functionally important fragments of heat shock protein (hHSP70), human tumor necrosis factor (hTNF) and protein $gp181$ from phage φKZ were obtained. The proposed approach can be used for *de novo* protein design.

1 Introduction

Proteins are an essential part of biological systems and processes. They have a number of unique properties, in particular, the ability to form a stable native 3D structure. We believe that there is a complex hierarchical structure in proteins that helps proteins to complete the folding process in a short time. Classical concept of the structural organization of proteins includes the following levels of organization:



In this hierarchy, only the whole proteins and structural domains have stable spatial organization. Usually, isolated elements of protein secondary structure do not keep the same shape as in the structure of the whole protein. Thus, structural domains are minimal objects that have the property of self-organization. Structural domain size varies from a few dozen to several hundred amino acid residues($N \approx 10^2$).

If an average time of transition between conformational states is 10^{-12} seconds and there are three major conformations of the polypeptide backbone, a protein folding time can be theoretically estimated as

$$T = 3^N * 10^{-12} \tag{1}$$

This exceeds the lifetime of the universe. This fact is called Levinthal paradox and was first given in [1]. Thus, existing views on the structural

organization of proteins do not allow us to give an adequate explanation of the protein organization. They do not let one see the real multilevel hierarchy in the structure of proteins and do not allow for the effective application of experimental methods. Thus, we needed new methods for structural organization of protein molecules identification and analysis. We supposed that these theoretical methods must be based on analysis of the protein sequences, because in early 60s it was suggested that "it may be concluded that the information for .... the assumption of the native secondary and tertiary structures, is contained in the amino acid sequence itself" [2].

To achieve this goal it was necessary:

- To prove the possibility of applying statistical methods to describe physical and chemical characteristics of the amino acid residues;

- To analyze the positional information entropy of natural protein sequences and identify features of the organization of information recorded in protein sequences;

- To develop a method for the protein sequences analysis, allowing us to identify the hierarchical structure of protein sequence;

- To carry out an experimental verification of the ANIS method;

- To develop new approaches for identification of functionally important parts in proteins and protein design.

2 Description of the physical and chemical characteristics of amino acid residues by statistical parameters

Multiple attempts to identify the laws in the arrangement of amino acid residues in protein sequence did not give significant results [3, 4, 5]. In particular, in a paper [5] it was shown that only 1% of residues in the protein sequence are non-random. This contains a contradiction: properties of the protein are determined by the physico-chemical characteristics of amino acids along the polypeptide chain, but amino acid residues are arranged in sequence almost randomly.

Apparently, it is necessary to use a new model of the structural organization of sequences based on statistical description of physical and chemical properties of amino acid residues to resolve this contradiction. The model proposed in work [6] considers environment of amino acid residues by sequence not individual residue. The result was an adequate description of the physicochemical parameters of amino acid residues (Fig. 1). Groups

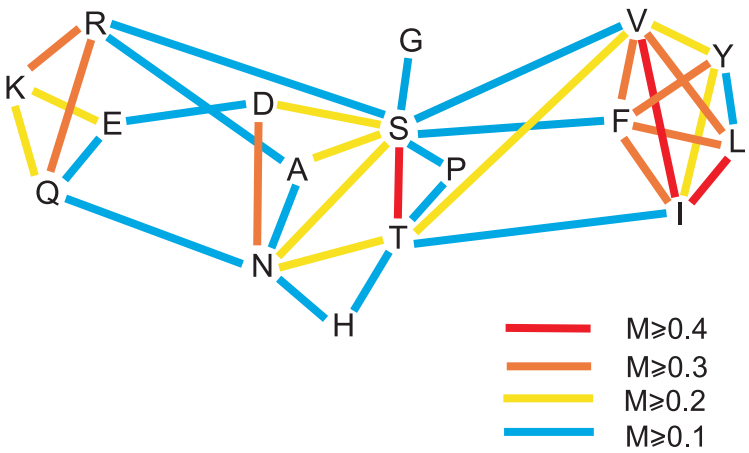


Figure 1: Similarity graph of amino acid residues based on the statistical model. M factor reflects the similarity between the amino acid residues. Greater coefficient correspond to a more similar residues. Red line connects residues with the similarity of $M \geq 0.4$, orange line connects residues with $0.4 > M \geq 0.3$, the yellow line connects residues with $0.3 > M \geq 0.2$, the blue line connects residues with $0.2 > M \geq 0.1$. The graph lacking residues C, M and W with a similarity coefficient $M < 0.1$ with any other residue.

of amino acid residues with different properties were identified using this approach: hydrophobic amino acid residues (L, I, V, F, Y), branched polar side chains (K, R, E, Q) and a group of residues with small side chains (D, A, N, T, S). The amino acid residues which contain functional groups with unique properties (C, G, H, M, P, W) have a unique environment in the protein sequence. It was shown that one needs to take into account not only isolated residues but residues and their environment in the protein sequence to describe the physical and chemical properties of the protein. Suggesting that a unit of the protein sequence is not only single amino acid residue, we identified a hierarchical organization in protein sequences.

3 The theoretical basis of the method

Initially we investigated the organization of protein sequence information to find the minimal size of the protein sequence unit. Different releases of NRDB database were used as the starting data. Releases differ significantly both in the number of protein sequences, and also consist of non overlapping data sets. Thus, we assumed that data sets in different releases are completely different [7, 8]. Based on these datasets of protein sequences, probability (P^k) of occurrence of various pairs of amino acid residues at fixed distances k (number of amino acid residues between them) were calculated (Fig. 1).

All database sequences were used for P^k matrices $20X20$ (according to number of amino acid residue types) calculations. P^k matrices were calculated for $k = 0, ..., 40$. Information entropy was calculated for each matrix by Shannon's equation (1) [9]. Since the size of the selected database affects entropy value, we used the value of S^0 as a normalization factor in order to neutralize this effect. The obtained values of the normalized information entropy S^k/S^0 as a function of distance k were calculated for three NRDB database releases and are shown in Fig.2.

$$S^k = -\sum_{i=1}^{20}\sum_{j=1}^{20} P_{ij}^k log_2 P_{ij}^k \tag{2}$$

The dependencies obtained for all three databases have identical (S-shaped) character with corresponding local maxima and minima (Fig.2). This suggests that the curve reflects entropy of the natural polypeptide chains sequence in general but not the database. For values $k > 30$ the dependence of $S^k/S^0(k)$ goes on plateau, i.e. there is no correlation between amino acid residues in the sequence on large distance k. This result

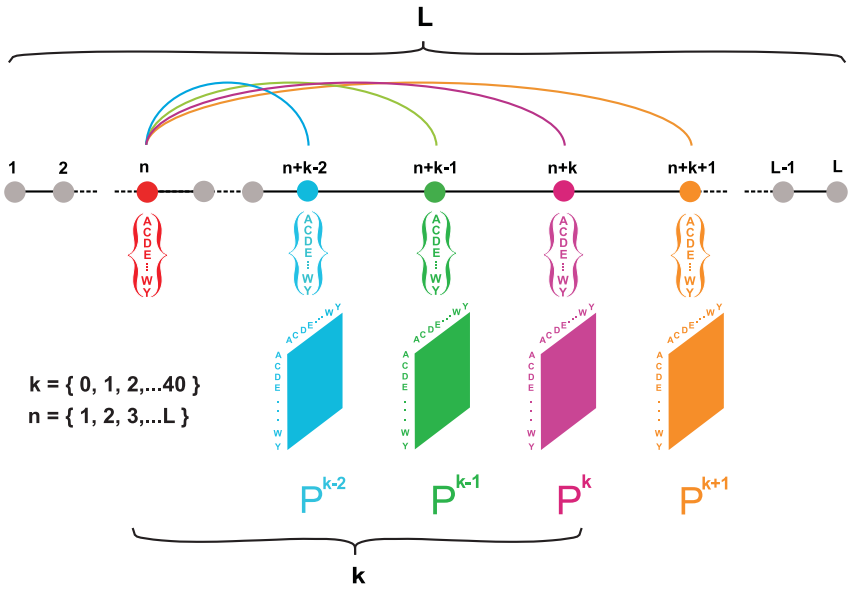


Figure 2: Computation scheme of amino acid residues occurrence frequencies matrices.

indicates that the stable structural elements of natural polypeptide chains should have characteristic size of 60 amino acid residues. This finding is well consistent with the lower limit of the structural domains from experimental data. Furthermore, it is necessary to note that an amplitude of normalized information entropy $S^k/S^0(k)$ fluctuations decreases with the rise of k.

The most interesting feature of the obtained dependence is a strong "jump" at $k = 5$ for normalized information entropy values. This fact implies that inside the fragment of five residues long there is the highest level of correlation between residues. So, this is the reason to assume that the most appropriate unit of natural polypeptide chains sequence description is pentapeptide.

Below we use the term "informational unit" to refer to fragment of five residues long.

4 The biological significance of oscillations

Visible oscillations (Fig.2) indicate the existence of correlations between amino acid residues at small distances. Fourier analysis revealed two pe-

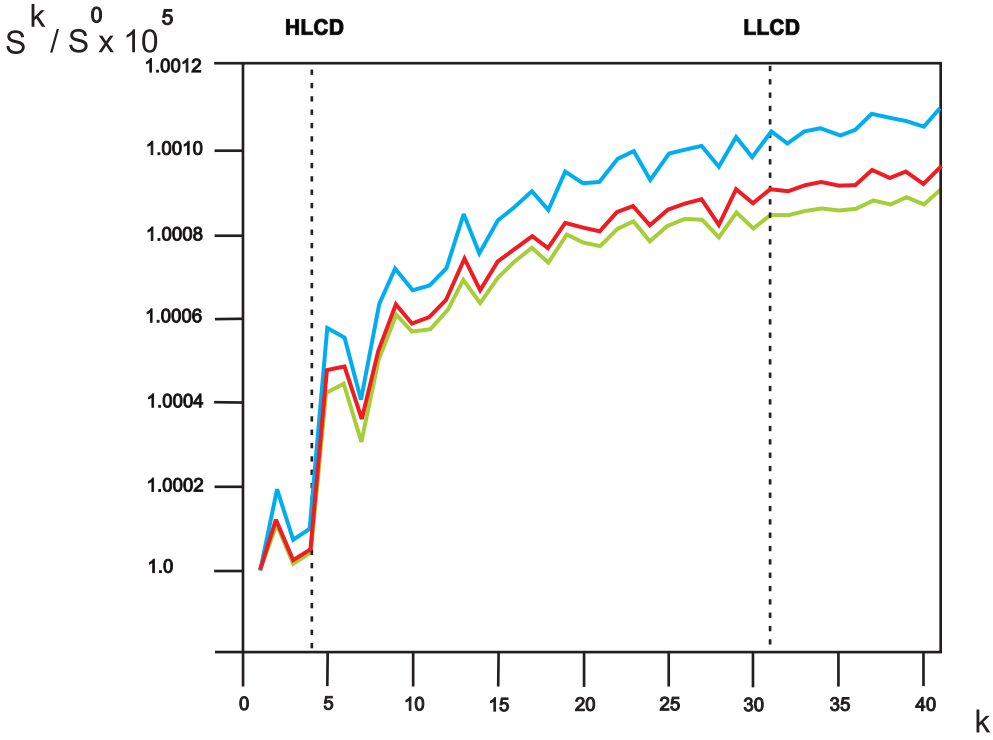


Figure 3: Dependence of the normalized information entropy S^k/S^0 from the distance between amino acid residues k.

riods, 3.6 and 2.9 amino acid residues. These periods correspond to the periodicity of α-helices and helix 3_{10}. These oscillations clearly visible in Fig.2 appear due to very high content of spiral fragment in natural proteins, with an average above 40%. Removing the oscillatory component by inverse Fourier transformation allowed us to restore a S-shaped curve [6].

5 Analysis of the Protein Sequence Information Structure

The method is based on the idea that short fragments of polypeptide chain - "information units"(IU) can be used for description of protein sequence. The sequence of proteins is regarded as a system of overlapping IU [10].

Let protein sequence $P = A_i$ be a sequence of amino acid residues $A_i, i = 1, 2, ...L$, where L is a length of protein sequence, and amino acid residues can be of 20 types. Let us consider some nonredundant protein sequence database(NRDB). Each subsequence $S = S_1, ...S_M$ of $M = 5$ amino acid residues has frequency $f(S)$ of occurrence in all sequences from the database NRDB. Now we choose a set of subsequences S' of 5 residues long, different from S by only one residue. We designate all such subsequences as "equivalent" IU. Each subsequence S' has frequency of occurrence $f(S')$ in all sequences from the database NRDB. The frequency function of occurrence for equivalent IU is a sum of corresponding frequencies $f(S')$

$$F(S) = f(s) + \sum_{S'} f(S'). \tag{3}$$

Let us take one protein sequence $P = \{A_i\}$ of L residues long. We will consider all possible overlapping subsequences S of M residues long. If $M = 5$ one can find $L - M + 1 = L - 4$ subsequences. Let us introduce the numbering of these subsequences $S = S_i$ of 5 residues long by their's central residues i, so, $i = 3, ..., L - 2$. Now we introduce function of "population" for each position in the sequence by equivalent IU:

$$F(i) = \sum_{i-2}^{i+2} F(S_i), \tag{4}$$

i.e. the sum of the frequencies of occurrence for the five information units, in which the amino acid residue i occurs in the protein sequence. So, for the protein sequence of L residues long "population" function $F(i)$ of equivalent IU (3) was constructed for each position i in the protein sequence. We now

pass from a discrete function $F(i)$ to a continuous function $F(x)$ representing the histogram $F(i)$. Further nonlinear smoothing was performed for $F(x)$ (Figure 3). We constructed a nonlinear smoothing function $G(a, x)$ as follows. Consider smoothing function $\varphi(x)$ - continuous function supported in the segment $[-1/2, 1/2]$, $\varphi(-1/2) = \varphi(1/2) = 0$, $\varphi(0) = 1$, $\varphi(x)$ takes positive values in the range $(-1/2, 1/2)$,increases monotonically in $[-1/2, 0]$, decreases in $[0, 1/2]$. The graph of the function is symmetric with respect to the straight $x = 0$. We assume that the function $\varphi(x)$ is smooth, the derivative is not equal to zero in the segments $(-1/2, 0)$ and $(0, 1/2)$. Thus, as the smoothing function, one can select the centrosymmetric positively defined function in some interval. In our work, we used an overstretched and shifted Gaussian function defined in a bounded interval.

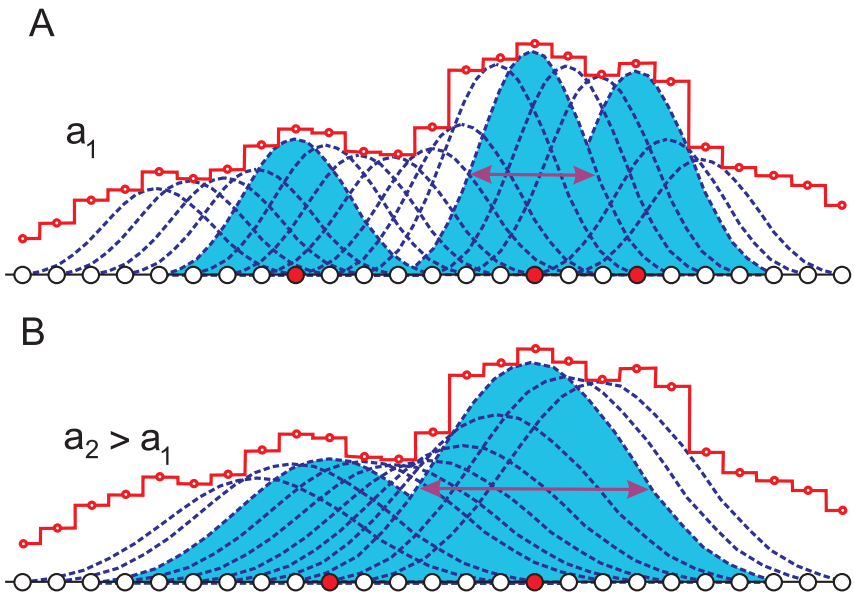


Figure 4: Nonlinear smoothing of "population" of equivalent IU function.

Now let consider shifts and stretching for smoothing function

$$\varphi^{(a,i)}(x) = \varphi(\frac{x-i}{a}) \tag{5}$$

when $a \geq 1$. Function $\varphi^{(a,i)}$ has support on the interval $[-1/2a+ia, 1/2a+ia]$. Let us define nonlinear smoothing function $G(x, a)$ for discrete function $F(i)$ as

$$G(i, a) = sup\ c, \quad c: \ c\varphi^{(a,i)}(x) \leq F(x), \forall x. \tag{6}$$

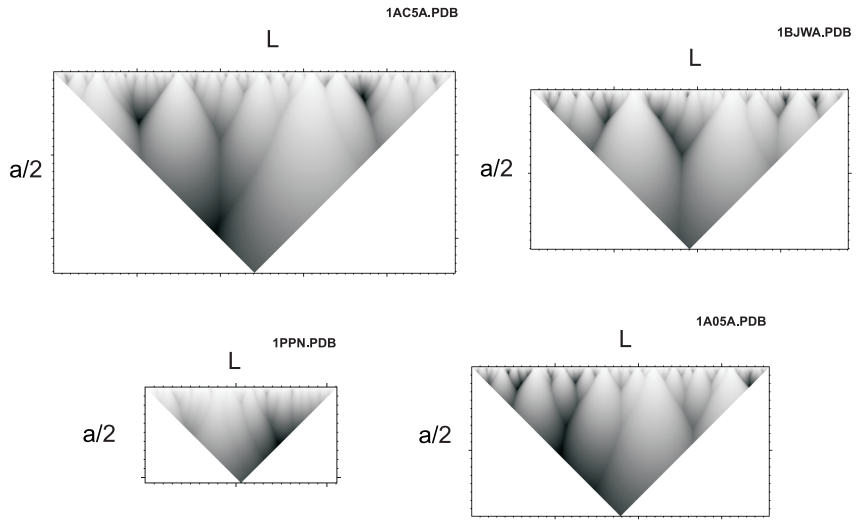


Figure 5: Examples of calculations of information structures for different proteins. The axes represent the numbers of residues of a sequence L and smoothing coefficient a. Darker regions correspond to the centers of regions of length a with a maximum level of coordination between information units in the protein sequence. The figure shows hierarchically organized structures.

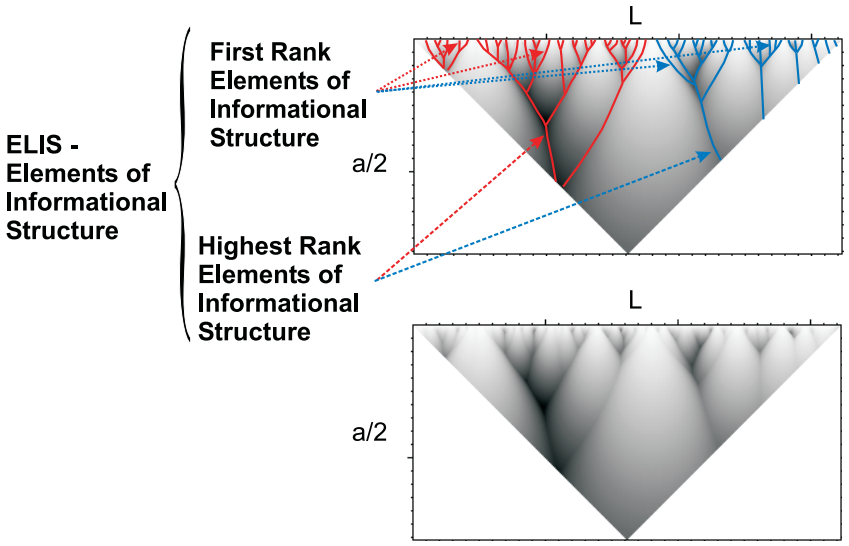


Figure 6: Information structure graph with marked the elements of information structures (ELIS) at different hierarchical levels.

Parameter a is called a scale of smoothing. The function $G(x, a)$ is nonzero in the interval $a \in [1, L-4], x \in [2+a/2, L-2-a/2]$. Function $G(x, a)$ has the support as triangle area on the plane with coordinates (x, a), vertices $(L/2, L-4), (2+1/2, 1), (L-2-1/2, 1)$ (Fig. 4). The whole set of values of the smoothing function $G(x, a)$ is called the information structure of the protein. Now let us mark all maxima of the function $G(x, a)$. If we replace the merge point lines of local maxima by vertices of the graph, and their connecting lines of local maxima by edges, then we obtain graph of the information structure (Fig. 5).

6 Information Structure Analysis as a method of protein engineering

In our opinion, currently only the described method of protein sequence analysis (ANIS method) allows to reveal a system of hierarchically organized elements in the protein sequence. The hierarchically organized elements were compared with the structural elements of proteins [10]. It was shown, that structural domains with clearly defined domain structure correspond to high rank ELISs (Figure.7). It should be noted that this cor-

relation is not affected by the type of secondary structure elements forming protein structural domains.

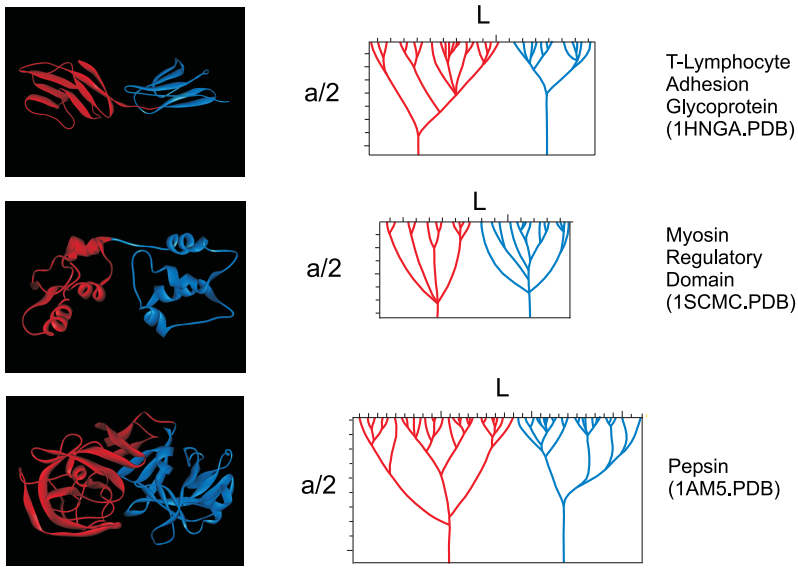


Figure 7: Protein structural domains correspond to high rank ELISs. Information structures of proteins with colored high rank ELISs are shown on the right. L is the number of amino acid residue in protein sequence. a is a scale of smoothing. Green dotted line limits the function $G(x, a)$ domain. 3D structures of the proteins are shown on the left. Color of the structural domain corresponds to color of the high rank ELIS.

It is known that structural domains are stable structural elements, and may independently form the 3D structure. Correspondence between the highest rank ELIS to structural domains may point to the important role of high rank ELIS in the formation of the spatial structure of proteins. We assume that the elements of 3D structure corresponding to the elements of information structure of other ranks are also stable elements of proteins 3D structure. This assumption allowed us to use of the ANIS method for protein design [11, 12]. Below are examples of the application of this method for identification of functional fragments with a specific feature in natural polypeptide chains.

6.1 Identification of the functional domain in protein *gp*181 from bacteriophage φKz

Some proteins of bacteriophage φKZ Pseudomonas aeruginosa are of especial interest. In particular, protein *gp*181 from base plate of the bacteriophage tail serves for local destruction of the Gram-negative bacteria cell wall and can be used to control the growth of microorganisms. A protein domain responsible for the enzymatic activity is a small part of a large protein *gp*181 of 2237 amino acid residues. The problem was to find the fragment of protein gp181 with desirable catalytic activity, capable for independent folding.

Information structure of the protein *gp*181 has several high rank ELIS (Figure 8). These protein fragments were produced as recombinant proteins [13]. Some of them (181 6−181 9) demonstrated desirable catalytic activity. The enzymatic activity of some protein fragments was 12 times higher then of lysozyme from chicken eggs, which was used for control.

So, application of ANIS method to the sequence of the *gp*181 protein allowed us to identify the protein fragment carrying desired catalytic activity.

6.2 Identification of the functionally important fragments in human tumor necrosis factor *h*TNF

In this study the ANIS method was used to identify functionally important fragments in the human tumor necrosis factor *h*TNF. *h*TNF is a multifunctional cytokine involved in the regulation of important physiological processes (anti-tumor immunity, immune cell proliferation, apoptosis, and others). *h*TNF is mainly synthesized by activated macrophages, T-lymphocytes and natural killer cells. *h*TNF is one of the major mediators of inflammatory processes in the human body. In clinical practice, recombinant monoclonal antibodies as well as fusion protein with extracellular domain of the receptor *h*TNF are used to neutralize *h*TNF molecules. However, the use of these drugs may be accompanied by dangerous side effects. Therefore it is important to develop *h*TNF antagonists. *h*TNF monomer is a β-structural protein. Active form of *h*TNF is a trimer of identical subunits with a molecular mass of 17 kDa (Fig. 9A). It was suggested that a protein region responsible for binding to *h*TNF receptor can act as *h*TNF antagonist. Previously it was shown that high rank ELISs correspond to stable elements of the spatial organization. So, it was decided to perform *h*TNF fragmentation on the basis of its information structure (Fig. 10). Peptides $D1$, $D2$, $D3$ and $D4$ (Fig. 11) corresponding to the highest rank

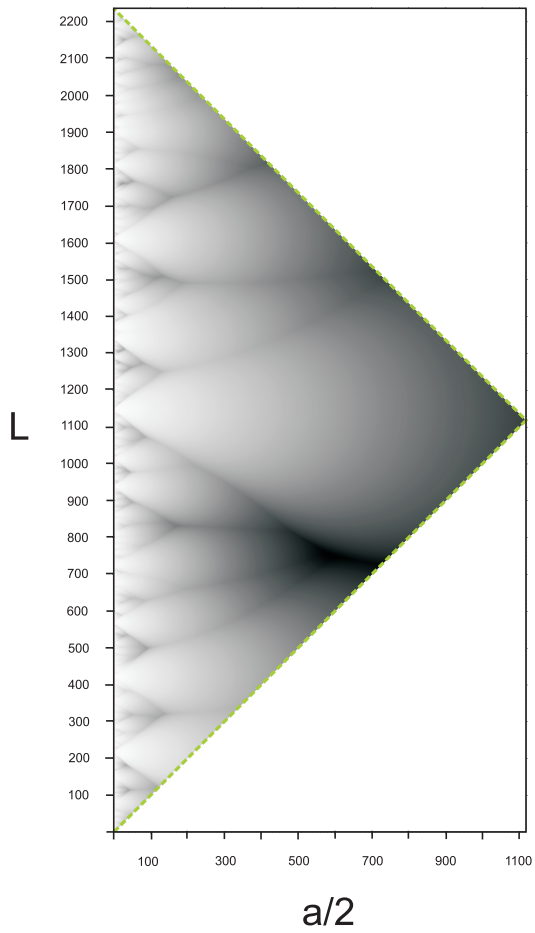


Figure 8: Information structure of the protein gp181 from bacteriophage φKz Pseudomonas aeruginosa. L is the number of amino acid residue in protein sequence. a is a scale of smoothing. Green dotted line limits the function $G(x, a)$ domain.

individual ELISs of hTNF (Figure 10) were obtained using the Escherichia coli expression system.

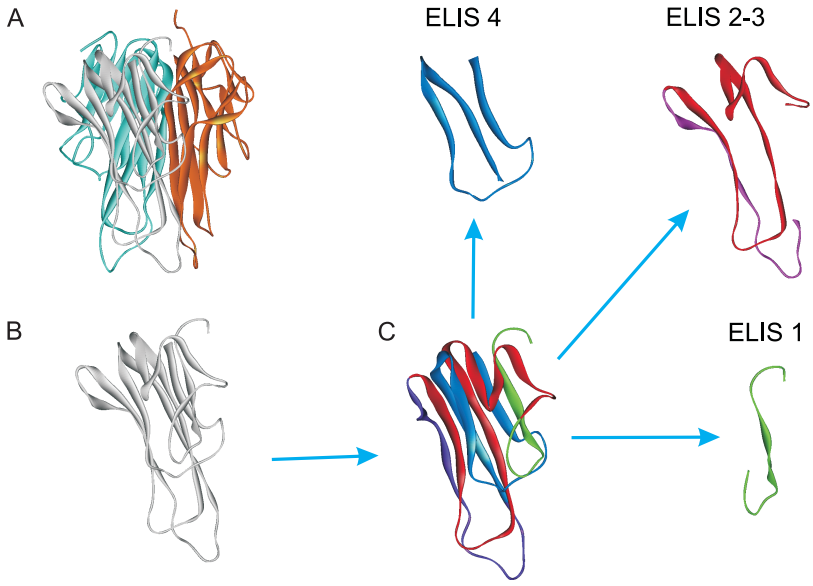


Figure 9: 3D structure of hTNF protein. ( ) Protein hTNF functions as a complex of three identical molecules. ( ) 3D structure of single hTNF protein chain. ( ) Single protein hTNF chain with specified fragments according to color coding of corresponding high rank ELIS (Figure 10). These fragments were obtained in an isolated form.

Peptides D1, D2, D3 and D4 (Fig. 10) contain some parts involved in the formation of β-structural elements (Fig. 9,C). So, producing them as separate peptides could lead to the formation of macromolecular aggregates. Some additives were used to prevent peptide aggregation. However, we failed to produce peptide D2 (Fig. 9).

Experiments have shown (Fig. 12) that the hTNF cytotoxic effect depends on the concentration of peptide D1. So, an excess of the peptide D1 reduces cytotoxic effects of hTNF. Thus, peptide D1 exhibits the desired properties of hTNF antagonist [14].

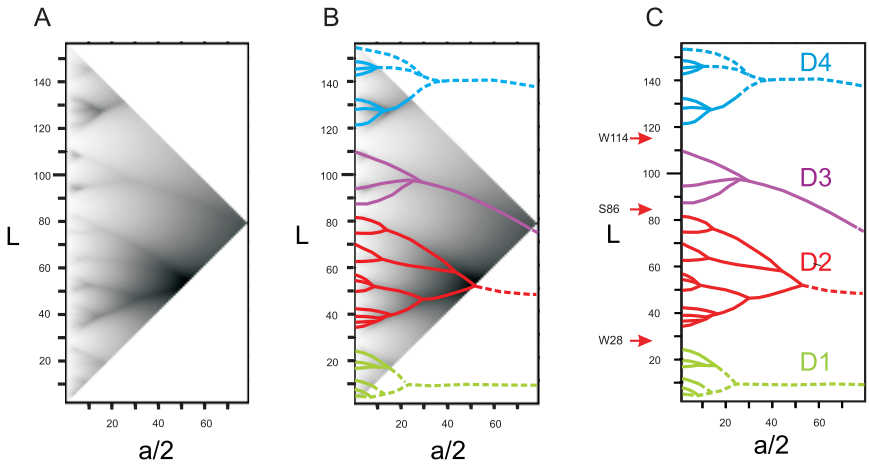


Figure 10: Information structure of the protein hTNF.L is the number of amino acid residue in the protein sequence. a is a scale of smoothing. Green dotted line limits the function $G(x, a)$ domain. B. Separated high rank ELISs are shown by different colors. C. Boundaries between different high rank ELISs are marked with an arrows.

D1 3 SSSRTPSDKPVAHVVANPQAEGQLQWLN 30

D2 31 RRANALLANGVELRDNQLVVPSEGLYLIYSQVLFKGQGCPSTHVLLTHTISRIAV 85

D3 86 SYQTKVNLLSAIKSPCQRETPEGAEAKPW 114

D4 115 YEPIYLGGVFQLEKGDRLSAEINRPDYLDFAESGQVYFGIIAL 157

Figure 11: Sequences of the D1, D2, D3 and D4 fragments of hTNF protein. The fragments correspond to different high rank ELISs (Figure 10).

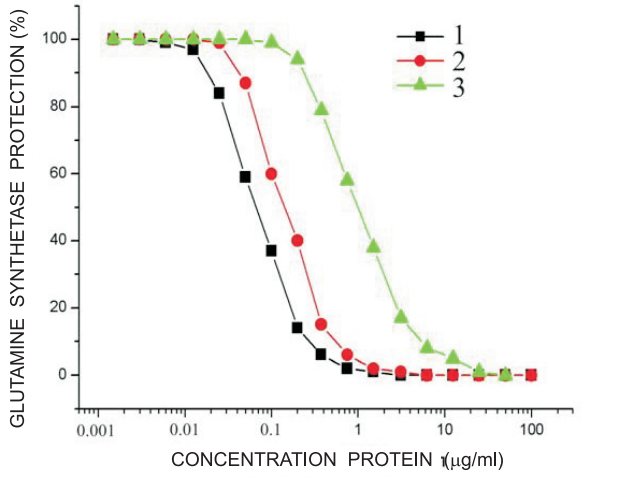


Figure 12: Cytotoxic effect on cells L929 of *h*TNF (1) and *h*TNF mixture with peptide D1 at a ratio 1 : 2 (2), 1 : 50 (3)

6.3 Identification of the functionally important fragments in the heat shock protein *h*HSP70 sequence

The main criterion for functionally important fragments identification described above was the correlation to high rank ELISs. However, sometimes fragments of the proteins are sufficiently large. It was proposed to use as an additional criterion the density of first rank ELISs distribution in polypeptide chain.

In the paper [15] it was shown that fragments of the polypeptide chains whith low density of the first rank ELISs distribution (ADD- sites) have the ability to form effective interaction through adaptive conformational rearrangements. This hypothesis was tested with a human heat shock protein (*h*HSP70).

One of the promising methods for therapy of various diseases is the activation of different factors of the innate immune system, including natural killer cells (NK-cells). NK-cells are a special population of lymphocytes from the innate immune system, that play an important role in antitumor and antiviral immunity. From the literature it was known that the activation of NK-cells by heat shock protein *h*HSP70 increases the production of γ-interferon (INF-γ).

The amino acid sequence of *h*HSP70 consists of 641 residues. The prob-

lem was to find a short peptide from *h*HSP70 with an activation effect on NK-cells for further use in clinical practice.

Preliminary studies have shown that the substrate-binding domain of *h*HSP70 makes desired stimulating effect on the production of γ-interferon by NK-cells. However, its length is 123 residues (according to high rank ELIS $G426 - M549$, Fig.13) and it is necessary to find a shorter fragment of *h*HSP70 for NK-cells activation.

Interaction site in the substrate-binding domain of *h*HSP70 is unknown. The following fragments of different information types (Fig. 14) were chosen for synthesis and experimental validation from the substrate-binding domain of *h*HSP70:

- Fragments 399-408, 411-424, 461-470 and 509-515 are ADD+ sites;
- Fragment 526-543 is the only one ADD- site in the substrate binding domain sequence of *h*HSP70.

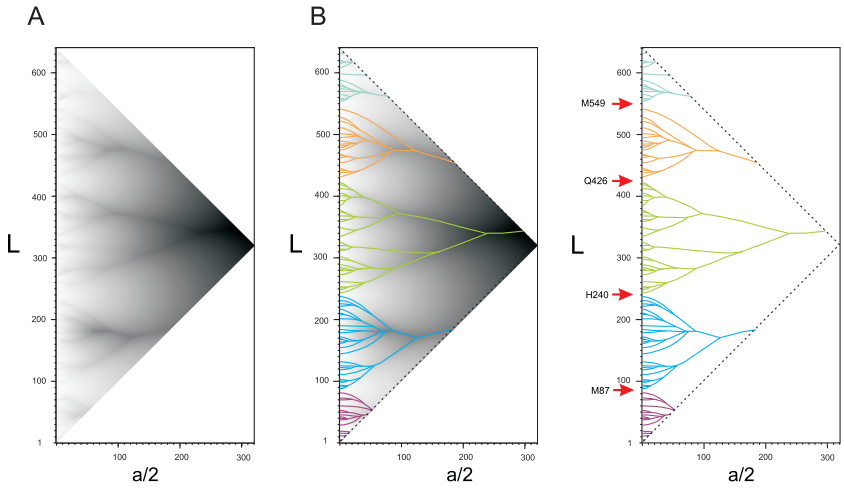


Figure 13: A. Information structure of the protein *h*Hsp70. B. Separated high rank ELISs are shown in different colors. L is the number of amino acid residue in the protein sequence. a is a scale of smoothing. Green dotted line limits the function $G(x, a)$ domain. Boundaries between different high rank ELISs marked with an arrows.

Also the fragment $450 - 463$ (TKD-peptide) was synthesized for functional activity testing. Biological activity of the TKD-peptide is known from the literature. For all synthesized peptides their effects on production of INF-γ by NK-cells and cytotoxicity were tested. The results are shown in Fig.15. One can see that INF-γ production with 526-543 peptide (of

399 408
LSLGLETAGG

411 424
TALIKRNSTIPTKQ

450 463
TKDNNLLGRFELSG

461 470
LSGIPPAPRG

509 515
RLSKEEI

526 543
KAEDEVQRERVSAKNALE

Figure 14: The fragments of different information types from high rank ELIS $G426 - M549$ of hHsp70. Fragments 399-408, 411-424, 461-470 and 509-515 are $ADD+$ sites. Fragment 526-543 is the $ADD-$ site in the substrate binding domain sequence of hHSP70. Fragment $450 - 463$ is the TKD-peptide with known functional activity.

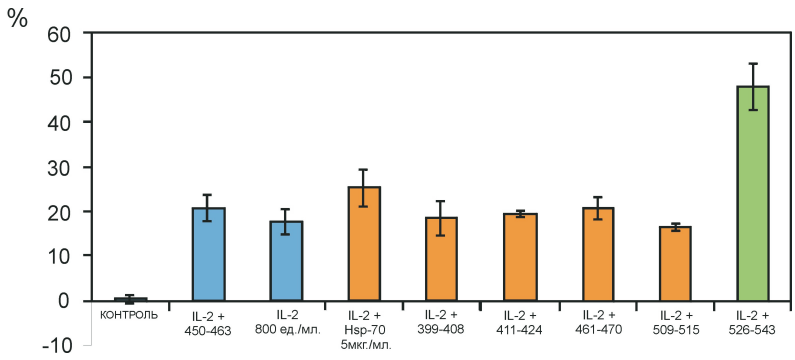


Figure 15: Effect of peptide fragments of hHSP70 on the production of $\gamma-$IFN by NK-cells shown by flow cytometry method. NK-cells were incubated with IL-2, hHSP70 and peptides (2 mg/ml) for 18 hours. Control is the unstimulated NK-cells.

ADD- type) twice higher then for 309-408, 411-424, 450-463, 461-470 and 509-515 peptides (ADD+ type). Thus, the effect of stimulating of INF-γ production by NK-cells was obtained only for the peptide of ADD- information type which have ability to adaptive conformational rearrangements. The results confirmed that the ability of polypeptide chain to adaptive conformational rearrangements ensures the formation of effective interactions between polypeptide chains [16, 17, 18].

7 Discussion

The developed method based on the idea that basic unit of protein sequences information is fragment of five residues. The proposed new approach of encoding protein sequences allowed to develop method of ANalysis of Information Structures of proteins (ANIS method). This method reveals a hierarchical organization of information structure in protein sequence. It was shown that protein information structure consist of hierarchically organized elements (ELISs), and correspond to stable elements of 3D structure of proteins.

The ANIS method was proposed:

- For the design of new recombinant proteins;

- For the isolation of fragments of proteins responsible for their functional activity.

New patterns of formation of effective interactions between polypeptide chains were identified. An important role of adaptive conformational rearrangements in formation of effective interactions between polypeptide chains was shown for enzyme-inhibitor complexes.

Acknowledgements

The authors thank Alexei Adzhubei for proofreading and valuable comments, Lada Petrovskaya, Viktoria Toporova and Elena Kovalenko for experimental proof of the method. The research was partially supported by Presidium of Russian Academy of Science Programm “Molecular and Cell Biology”, grant $14 - 04 - 90034$Bel_a of Russian Foundation for Basic Research and the Program of Fundamental Research of the Presidium of the RAS within the strategic directions of science development for 2014 “Fundamental problems of mathematical modeling” and the project “The mathematical model of spatial organization of natural polypeptide chains on the base of the information content of primary structure”.

References

[1] C. Levinthal, How to fold graciously. *In Mossbauer Spectroscopy in Biological Systems, Proceedings of a Meeting held at Allerton House*, Monticello, Illinois (1969).

[2] C.B. Anfinsen, E. Haber, M. Sela, F.H. White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc.Natl.Acad.Sci. USA* **47** (1961), 1309-1314.

[3] O.B. Ptitsyn and M.V. Volkenstein. *Protein structure and neutral theory of evolution.* J.Biomol.Struct.Dyn., 4 (1986), 137-156.

[4] O. Weiss, M.A. Jimenez-Montano, H. .Herzel "Information content of protein sequences" J. Math. Biol. (2000) 206:379-386.

[5] G. Szoniec, M.J. Ogorzalek, *Entropy of never born protein sequences.* SpringerPlus, 2013.

[6] S.I. Rogov, A.N. Nekrasov "A numerical measure of amino acid residues similarity based on the analysis of their surroundings in natural protein sequences" Protein Eng. (2001) vol. 14 (7), pp.459- 463 .

[7] L. Holm, C. Sander, Removing near-neighbour redundancy from large protein sequence collections.*Bioinformatics* **14.5** (1998), 423-429.

[8] W. Li, L. Jaroszewski, A. Godzik, Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* **18.1** (2002), 77-82.

[9] C.E. Shannon, A Mathematical Theory of Communication. *Bell System Technical Journal* **27** (1948), 379-423.

[10] A.N. Nekrasov, Analysis of the information structure of protein sequences: a new method for analyzing the domain organization of proteins. *J. Biomol. Struct. Dyn.* **21(5)** (2004), 615-624.

[11] A.N. Nekrasov, V.V. Radchenko, T.M. Shuvaeva, V.I. Novoselov, E.E. Fesenko, V.M. Lipkin "The Novel Approach to the Protein Design: Active Truncated Forms of Human 1-CYS Peroxiredoxin" J. Biomol. Struct. Dyn. (2007) vol. 24(5), pp.455-462,

[12] Nekrasov A.N., Petrovskaya L.E., Toporova V.A., Kryukova E.A., Rodina A.V., Moskaleva E.Y., Kirpichnikov M.P. "Design of a novel interleukin-13 antagonist from analysis of informational structure" Biochemistry (Mosc). 2009 vol. 74(4), pp. 399-405

[13] Yves Briers, Konstantin Miroshnikov, Oleg Chertkov, Alexei Nekrasov, Vadim Mesyanzhinov, Guido Volckaert, Rob Lavigne "The structural peptidoglycan hydrolase gp181 of bacteriophage KZ" Biochem. Biophys. Res. Commun. (2008) vol. 374(4), pp. 747-751.

[14] Shingarova L.N., Petrovskaia L.E., Nekrasov A.N., Kriukova E.A., Boldyreva E.F., Iakimov S.A., Gur'ianova S.V., Dolgikh D.A., Kirpichnikov M.P. "Production and properties of human tumor necrosis factor peptide fragments" Bioorg Khim. 2010 vol. 36(3), pp. 327-336 Russian.

[15] [A.N. Nekrasov, A. . Zinchenko "Structural Features of the Interfaces in Enzyme-Inhibitor Complexes" J. Biomol. Struct. Dyn. (2010) vol. 28(1), pp. 85-96.

[16] L.M. Kanevski, P. Vlaskin, L. Alekseeva, A. Nekrasov, Y. Strelnikova, E. Kovalenko "Heat shock protein 70 and its peptide fragments increase IFN-gamma production and modify surface marker expression in human natural killer cells" 2th European Congress of Immunology, Berlin, Germany. European Journal of Immunology, 2009, Supplement 1, p. S343.

[17] Kovalenko E.I., Kanevskiy L.M., Sapozhnikov A.M., Nekrasov A.N. ."Application of the method of analysis of protein structure information to identify areas of heat shock protein 70 activating NK-cells" / In "Stochastic and computer modeling of systems and processes" Hrodno, 2011, ISBN-978-985-515-495-3 pp. 379-384 Russian.

[18] Kanevskiy L.M., Vlaskin P.A., Alekseeva L.G., Nekrasov A.N., Sapozhnikov A.M., Kovalenko E.I. "The fragments of heat shock protein 70 enhances the production of -interferon and affect the expression of surface markers on NK-cells" Medical immunology (2007) vol. 9(2-3), p. 142 Russian.

Cognition as a Dynamical Process: Mathematical Modeling

Eva Čadež [a]
Cognitive & Information Sciences, University of California, Merced, 5200 North Lake Road, Merced, CA 95343, USA

Michael J. Spivey [b]
Cognitive & Information Sciences, University of California, Merced, 5200 North Lake Road, Merced, CA 95343, USA

Vladimir M. Čadež [c]
Astronomical Observatory, 11160, Belgrade, Serbia

Abstract

Dynamical systems are those that change over time and they are widely present in nature. In this contribution, a typical radiative transfer equation for a physical process of radiation propagating through an interactive non-uniform medium is applied to an abstract-level cognitive science to model a dynamical process of learning and forgetting of episodes. The aim of this work is to use the introduced model to provide an insight into possible solutions for the dynamical system and shed light of significance of

[a] e-mail address: ecadez@ucmerced.edu
[b] e-mail address: spivey@ucmerced.edu
[c] e-mail address: vcadez@aob.bg.ac.rs

freely chosen parameters involved. Such a theoretical understanding can eventually point out the directions of future experimental work in the field of cognition including perception-action, memory, attention, learning, language, etc.

1 Introduction

In psychology and cognitive science, processes that evolve in time are predominantly modeled and analyzed by time series and linear regression mathematical/statistical methods. One reason for this that these provide one perspective on processes involved and may produce reasonable predictions about, say, what people will remember after learning some material and then engaging in some other activity for some shorter or longer time period. The other reason is that most psychologists and cognitive scientists are not mathematicians or statisticians so they are not trained in a huge variety of more complex methods of modeling. Cognitive concepts are mostly highly abstract hypothetical constructs and this necessarily introduces questions such as whether we can really measure them or interpret them in any scientific manner, for example. The entire era of behaviorism in psychology was devoted to pointing out problems like these [1]. A consequence, one of many, of this rigorous scientific criticism as well as of philosophical debates about methods of psychology is that a simplicity of models used is highly valued, perhaps to the point of oversimplification of studied phenomena. Along with many other influences during a development of psychology and cognitive science, these issues resulted in relatively limited variety of mathematical tools used in modeling. Many existing tools are completely adequate for some purposes but, we argue, even then, given the complexity of cognition, other windows of insight, namely, other mathematical tools should be popularized, mathematical dynamical systems in particular.

Cognition may be treated as a complex dynamical system and descriptive methods of investigating dynamics of systems have been increasingly present among tools cognitive scientists use (e.g., [2]-[4]). Qualitative fits are becoming more readily discussed. For example, Navarro, Pitt, & Myung [5] suggest that comparison of models based only on their quantitative fit to specific data (local model analysis) may be limiting our understanding of cognition. Capabilities of a model to fit various other possible data sets (global model analysis) is suggested in evaluating models of cognition as it may prevent oversimplified views of processes. Differential-equations-related modeling using analytical and numerical solutions to simulate pro-

cesses within a period of time during which the system changes , however, is usually not a part of an interdisciplinary training of cognitive scientists. We use this approach in our work and treat episode processing as a function of an "episode processor," a complex subsystem of a more complex dynamical system [6]. This work is a novel approach to episode processing in another sense as well – although the parts of this processing such as remembering ordered items or events, or attention in learning and memory are widely studied on their own elsewhere, they are integrated together here in a new way.

2 The model

The first parameter in the model is time. The complex dynamical system changes in time. Second, any information in the system may be more or less activated, so this intensity, I, is the second dimension of the system. Next, there is a dimension, the Conceptual Space (CS), along which concepts such as those labeled by words are distributed. If two concepts are closer together along this dimension they are more psychologically similar to each other (for more discussion on using continuous dimension, one-dimensional concept space, etc, see [6]). The order of events/concepts is an important feature of episodes and it is a separate dimension, independent of the CS. Simply, along this dimension the information activation is increased at places corresponding to position one, two, three, and so on. Cognitive linguistics researchers suggest that representation of time may be based on less abstract representation of space [7]. Following this research, we postulate that the order in time is treated in cognitive systems as the order in space and we call this dimension time-as-space dimension (TAS). Each activated information is represented as an activation of a specific position along each TAS and CS dimensions. We use Gaussians to represent this activation in order to acknowledge that a single concept also automatically activates some other close concepts and that the amount of this activation reduces with distance from the central concept. We use the same approach with the TAS dimension out of convenience. In experiments on word list memory, for example, concepts along the CS dimension are as distinct as positions in lists (TS), which we model by Gaussians of variance relatively small comparing to the distance between the means of those Gaussians. Finally, the TAS and CS dimensions are combined in the following way: the activation for position one, for example, is added to activation along entire CS dimension. When only certain concepts (information) are activated, only their activations increase the activation for the TAS dimension, so that

the resulting peaks of activation are able to represent which positions with TAS and CS coordinates were activated in the system at a time (Fig. 1). It is,finally, assumed that only the activations that reach a certain threshold get registered in another level of the processor and can be "read" later to represent an episode recall (altogether with the order of events/concepts). Everything else being equal, the more intense activation peak is both more likely to be recalled and faster to recall than a low intensity information.

The mathematical model has the following form for both CS and TAS dimensions.

$$\frac{dI}{dt} = -\alpha\,(t,x)\,I\,(t,x) + S\,(t,x) + C\,(t,t';x,x') \tag{1}$$

where:

$$C\,(t,t';x,x') = \int_{-\infty}^{\infty} dx' \int_{0}^{t} dt'\; W(t,t';x,x')\,I(t',x') \tag{2}$$

is the spatial and temporal correlation represented by the integration over x' and t'. The functional dependencies W are the weight functions determining influences of the temporal and spatial context at (t',x') on the selected item at (t,x). Their analytical expression has to be specified by modelers. For example, in reported simulations the model uses lateral inhibition at greater distances and excitation at smaller distances between items. In addition, small lateral activations at the x' and t' contribute less to the activation in x than greater activations (both negative and positive) while at the same time there is also a limit to the possible contribution of the large activations from other positions. This is achieved by using a Gaussian weighting function for distance contributions and a log function for intensity contributions from a site x' to the site x and from a point at t' to the point t.

Information intensity $I(t,x)$ spontaneously decays in time if no additional influences are present:

$$\frac{dI}{dt} = -\alpha\,(t,x)\,I\,(t,x) \tag{3}$$

with the analytical solution:

$$I\,(t,x) = I_0(x)\,\mathrm{e}^{-\tau(t,x)} \quad \text{with} \quad \tau\,(t,x) \equiv \int_{0}^{t} \alpha\,(t',x)\,dt' \tag{4}$$

The quantity $\tau(t,x)$ is positive, i.e. $\tau > 0$, and possibly can be considered as "a psychological time interval."

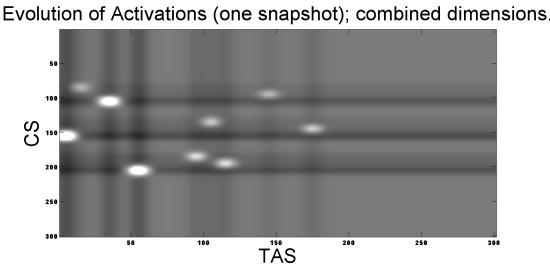


Figure 1: The TAS and CS dimensions are combined and the system changes in time. At each time point, the distribution of activations changes. The highest activations are represented as the lightest in this figure

This part of the equation describes an exponential decay of activation at the reducing rate. Obviously, many activations will decay only to some non-zero level. The reducing rate trend of the decay may be used to model consolidation phenomena in memory – after initial rapid forgetting, the remaining activations do not change much in a long period of time.

When simulations are run, a slice of the simulation at one time point may look like activation distribution in Fig. 1.

3 Discussion of simulations and applications

3.1 External and internal inputs into the system; learning and rehearsal

The term $S(t,x)$ in the Eq1 modifies the initial distributions both in time t and in CS or TAS dimensions. The process of "activation gain", requires $S(t,x) > 0$. Alternatively, if $S(t,x) < 0$, one can think in terms of a "activation drain." For example, if a person has some initial distribution of knowledge, it can be manipulated by practice, for example. This, of course, may result in more adequate knowledge but also, with bad training, may lead to misconceptions. In addition, the source of input may come from an internal rehearsal of information, for example. We label this $R(t,x)$ but it is equivalent to the positive $S(t,x)$. In this, too, the input to the system may be beneficial by adding activation to the active information. Alternatively, if this input is at the "wrong' location, it will be detrimental to knowledge. A simple variant of the general equation of the model could, in this case, be the following:

$$\frac{dI}{dt} = -\alpha(t,x)\, I(t,x) + S(t,x) + R(t,x) \tag{5}$$

3.2 Coupled memory systems or two person interactions

If two dimensions representing two kinds of systems such as knowledge of two people or information in two memory systems, for example, interact with each other, a system of two coupled differential equations can further model joint evolution of the two (or more, in principle) kinds of information a trace holds in an episode processor:

$$\frac{dI_1}{dt} = -\alpha_1(t,x)\, I_1(t,x) + \beta_1(t,x)\, I_2(t,x) + S_1(t,x) + C_1(t,t';x,x';t,x) \tag{6}$$

$$\frac{dI_2}{dt} = -\alpha_2(t,x)\, I_2(t,x) + \beta_2(t,x)\, I_1(t,x) + S_2(t,x) + C_2(t,t';x,x';t,x) \tag{7}$$

where, C_1 and C_2 have the form defined in Eq. 2. In principle, this kind of system has novel kinds of possible solutions, oscillations, for example.

3.3 False memory and attention

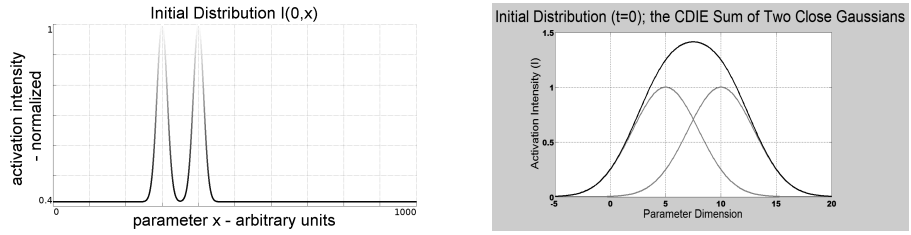


Figure 2: LEFT: The two far Gaussians. RIGHT: The two close Gaussians producing a false memory in between them.

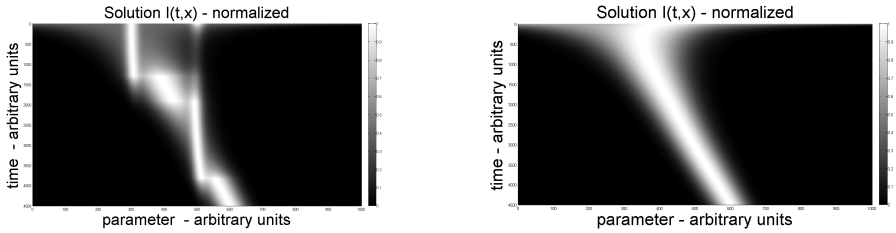


Figure 3: LEFT: Normalized solution of the development of the sum of the two far Gaussians. RIGHT: Normalized solution of the development of the sum of the two close Gaussians.

Consider two situations when only two Gaussians are present in the system. In Fig. 2, shown on the left are two concepts that are at such

a distance that they minimally influence each other. On the right are the Gaussians that are close: concepts activated here are very similar. A large peak is produced between them. In both cases, after activations are introduced to the system, they are integrated with each other and other activations and will evolve according to given equations. At some point in time the attenuation function, simulating attention, may be introduced in this system and the changes over time may be simulated (Fig. 3). The minimum of this (hyperbolic, in our case)function is to the left of both of the maxima, so that it asymmetrically influences them. The sign of the attenuation function in principle may be positive or negative, so that it will either add or reduce activations of information. In these figures, the time evolution "flows" from the top to the bottom. At early times, the two original high activations are visible, as well as the relatively inactive region between them. However, as the time passes, the attention adds activation to the left peak more than to the right one and the most strongly activated positions of the field change. This may represent the situation when a false memory appears due to the attentional process. Note that Figure 2, the right panel, shows another, different possibility of false memories formation – the two initial Gaussians are close enough to begin with, so that the area between them becomes highly activated forming a false memory at that position. In this case, this "false" peak may be higher than the original peaks and the attention may play a role in suppressing this "false" peak. This nicely simulates the results of warning subjects about the possibility of forming false memories when remembering lists of related words, so that they pay attention and suppress them.

4 Conclusion

In sum, in this short distribution we presented a dynamical model of several integrated processes in an episode processing and in brief we illustrated how simulations of complex systems may produce insights into complexities of integrated cognitive processes. Dynamical simulations may show some counter-intuitive behaviors and explanations because they take into account more interactions than a researcher can follow only by theorizing about phenomena, which is typically still the basis for the producing model assumptions in psychology and cognitive science.

References

[1] Skinner, B. F. (1977). Why I Am Not a Cognitive Psychologist, *Behaviorism, 5,* 2, pp. 1-10

[2] Grossberg, S. (1964). The theory of embedding fields with applications to psychology and neurophysiology. Rockefeller Institute for Medical Research.

[3] Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986). Schemata and sequential thought processes in parallel distributed processing. *J. L. McClelland, D. E. Rumelhart & the PDP Research Group, Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models.* Cambridge, MA: MIT Press/Bradford Books. 757.

[4] Van Geert, P. (1991). A dynamic systems model of cognitive and language growth. *Psychological Review, 98,* 3-53.

[5] Navarro, D. J., Pitt, M. A. & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47-84.

[6] Čadež, E. (2014). Integration of Memory, Perception, and Attention in Episode Processing. *Journal of Integrative Neuroscience, 13*, 1.

[7] Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12,*49-100.

Modeling CRISPR/Cas Regulation: Analysis of an Advanced Bacterial Immune System

Magdalena Djordjević [a]
Institute of Physics Belgrade, University of Belgrade, Serbia

Marko Djordjević [b]
Faculty of Biology, University of Belgrade, Studentski trg 16, Belgrade, Serbia

Abstract

Protection of bacterial cells against virus infection requires expression of molecules that are able to destroy the incoming foreign DNA. In CRISPR/Cas systems - which are recently discovered bacterial immune systems - this toxicity is (in part) avoided through rapid transition of the expression of the toxic molecules from "OFF" to "ON" state. Specifically, CRISPR array expression involves a mechanism where a small decrease of unprocessed RNAs leads to a rapid increase of processed small RNAs. Surprisingly, this rapid amplification crucially depends on fast non-

[a] e-mail address: magda@ipb.ac.rs
[b] e-mail address: dmarko@bio.bg.ac.rs

specific degradation of the unprocessed molecules by an unidentified nuclease, rather than on large cooperativity in protein binding. Furthermore, the major control elements that are responsible for fast transition of R-M and CRISPR/Cas systems from "OFF" to "ON" state, are also directly involved in increased stability of the steady states of these systems. We here discuss mechanisms that allow rapid transition of toxic molecules from unproductive to productive states in CRISPR/Cas systems.

1 Introduction

Bacterial immune systems defend host cell against infection by bacteriophages (bacterial viruses). A prominent examples of such system is the recently discovered CRISPR/Cas (Clustered Regularly Interspaced Short Palindromic Repeats/CRISPR *as*sociated sequences) systems. In order to defend host bacteria against the incoming infection, the immune systems have to express molecules that can destroy genome of the incoming virus. While these molecules are evidently useful, they can also be toxic, due to autoimmunity problems. That is, the same mechanism that is responsible for destruction of the foreign DNA, can also, in principle, lead to the destruction of the host genome.

An example of the balance between toxicity and usefulness is provided by the restriction enzyme within a Type II restriction modification system (R-M system) [1]. Since the restriction enzyme makes cuts in specific DNA sequences, it can, in principle, cut both the DNA of the incoming virus and host DNA. Destruction of the host DNA is prevented by methylase, which protects the same DNA sequences that are cut by the restriction enzyme. That is, while the restriction enzyme makes cuts in specific DNA sequences, methylase protects the same sequences that are being cut by the restriction enzyme. Consequently, unmethylated DNA sequences of the incoming virus will be cut by the restriction enzyme, while destruction of the host genome is prevented by its methylation.

A quite different type of bacterial immune system is provided by a recently discovered CRISPR/Cas system [2-3]. The system consists of CRISPR array and associated *cas* genes [4], and is represented by Fig. 1. CRISPR cassettes consist of identical direct repeats of about 30 bp in length, interspaced with variable spacers of similar length. CRISPR presents an adaptive prokaryotic immune system, which is responsible for defending prokaryotic cell against invaders, so that a match between a CRISPR spacer and invading phage (bacterial virus) sequence provides immunity to infection.

In addition to CRISPR cassettes, CRISPR-associated (*cas*) genes are also required for this immunity. Experiments show that the entire CRISPR locus is transcribed as a long transcript (called pre-crRNA) [5-6], which is further processed by *CasE* in *E. coli* to small interfering RNAs (called crRNAs) [6-7]; crRNAs are responsible for recognition and - together with a large complex that is formed by *Cas* proteins - inactivation of invading viruses [4].

From the above discussion, it is evident that bacterial immune systems can employ a quite different mechanisms for expression of toxic molecules. Despite these differences, it may also be useful to think in terms of more general principles that govern expression of toxic molecules inside bacterial cell. For exam-

ple, expression of a toxic molecule should generally be accompanied by expression of an antidote (e.g. methylation in the case of R-M systems). Furthermore, it seems plausible that generation of a toxic molecule should involve a rapid transition from "OFF" to "ON" state, so that toxic molecules are present in small amounts when they are not needed, but are then rapidly generated upon infection by invasive DNA. Finally, additional, more subtle, principles may also be relevant: e.g. fluctuations of the toxic molecule in its steady state might need to be small, in order to evade that a large fluctuation of the toxic molecule is unmatched by the antidote amount. We will below discuss relevant theoretical and experimental results on bacterial immune systems, with the purpose of pointing to some possible strategies for expression of toxic molecules inside cell.

Figure 1. A scheme of CRISPR/Cas genomic arrangement. Genomic arrangement of different *cas* genes and CRISPR array elements is indicated. R and S within the CRISPR array correspond, respectively to repeats and spacers; note that the spacer sequences differ from each other, and are labeled by consecutive numbers (1,2,3,..). IGLB and L in the figure correspond to the intergenic regions where promoters for, respectively, *cas* genes and CRISPR array are located. Different cas genes are labeled by cas1-3 and casABCDE.

2 A model of CRISPR transcript processing

In this section, we will analyze a mechanism for the fast transition from unproductive to productive state of the toxic molecule, which involves control at the level of transcript processing.

In *E. coli*, promoters that transcribe CRISPR cassettes and *cas* genes are distinct, and are (at least under normal growth conditions) considered to be poorly active due to repression by H-NS transcription factor [5]. While it is clear that CRISPR/Cas system in *E. coli* is functional [6, 8], virus infection in itself appears not to lead to system induction (at least under normal conditions) [9], and physiological conditions under which the system is induced yet have to be determined [4]. Consequently, functioning of this system has been investigated by either artificial overexpression of *cas* genes from plasmids, or by inhibition of H-NS repression of *cas* and CRISPR promoters [6-7, 10]. Surprisingly, quantitative measurements show that overexpression of *cas* genes leads to generation

of a very large number of crRNAs from only few pre-crRNAs [6]; specifically, disappearance of only a few pre-crRNA molecules normally present in the cell leads to a two orders of magnitude increase of crRNA upon *cas* overexpression.

We below summarize the main experimental observations, which we use to formulate a model of CRISPR transcript processing:

i) Endogenous (uninduced) levels of pre-crRNAs and crRNAs are low (~10 copies per cell) [6-7, 10], which was reported to be a consequence of repression of *cas* and (to a smaller extent) CRISPR promoters by H-NS [5].

ii) One of the Cas proteins (CasE) is responsible for processing pre-crRNAs to crRNAs [6-7]. When CasE is overexpressed, the amount of crRNAs increases for about two orders of magnitude, while the amount of pre-crRNAs drops to only few transcripts per cell [6]. Overexpression of CasE affects only the processing rate of pre-crRNA to crRNA, since it has been shown [6] that CasE does not influence either pre-crRNA transcription rate or crRNA stability.

iii) In addition to being processed by CasE, pre-crRNA is also degraded by an unspecified nuclease [5-6]. As a consequence of this degradation, pre-crRNA decays with a half-life of ~1 min without generating crRNAs. On the other hand, crRNAs are observed to be much more stable [6].

iv) It is currently unclear how CRISPR/Cas system is induced under natural conditions [4]. It was, however, shown that the repression of the *cas* promoter by H-NS can be relieved by a transcription activator (LeuO) [10]. It was consequently proposed that the endogenous system induction may involve activation of *cas* and (to a smaller extent) CRISPR promoters, through abolishment of H-NS repression [5].

The minimal model of CRISPR transcript processing, which is in accordance with the experimental observations summarized above, is schematically shown in Fig. 2. As can be seen from the scheme, the unprocessed transcripts (pre-crRNAs) are transcribed with certain rate, and are consequently either non-specifically degraded with rate λ_u, or processed by CasE to crRNA that are further degraded with rate λ_p. By non-specific degradation, we mean degradation that does not lead to accumulation of crRNA, and which is exhibited by an unidentified nuclease(s). All the parameters that characterize the system were either directly experimentally determined, or can be inferred from the experimental data by using the model described above [11]. In particular, for the discussion below, it is relevant that there is a very fast decay of pre-crRNAs and a

slow decay of crRNAs, so that the decay rates of pre-crRNAs and crRNAs are, respectively, $\lambda_u \sim 1$ min^{-1}, $\lambda_p \sim 1/100$ min^{-1}.

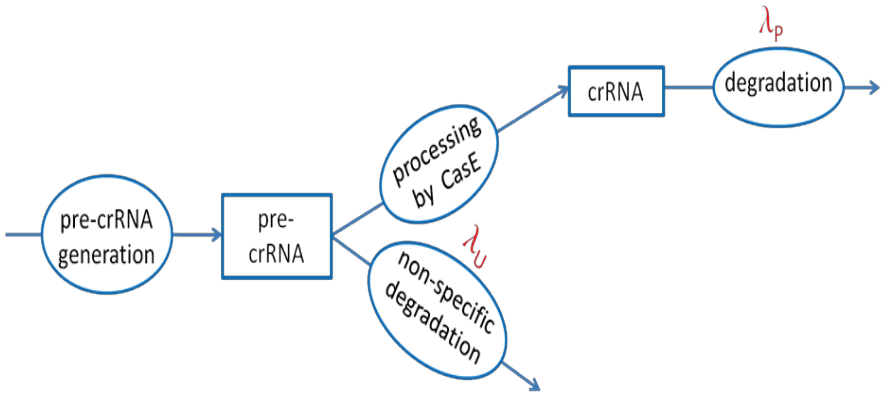


Figure 2. CRISPR transcript processing scheme. Pre-crRNAs are generated with certain rate, and are consequently either (non-specifically) degraded with rate rate λ_u, or are processed to crRNAs by CasE; generated crRNAs are then degraded with rate λ_p.

3 Large amplification of crRNA

We will first analyze the experimental observation that a small decrease of pre-crRNAs leads to a very large (two orders of magnitude) increase of crRNAs. It is evident that this large `amplification' of crRNAs is directly relevant for fast transition of the system from "OFF" to "ON" state. We will below denote concentrations of the unprocessed (pre-crRNA) and processed (crRNA) transcripts by, respectively, $[u]$ and $[p]$. Furthermore, we will use primes to indicate the quantity values after the system induction (e.g. after CasE overexpression). Consequently, changes in the number of pre-crRNAs and crRNAs are, respectively, labeled as $\Delta[p]=[p]'-[p]$ and $\Delta[u]=[u]'-[u]$.

Detailed kinetic equations that correspond to the model discussed above are provided in [11]. From these equations it is straightforward to derive the relationship between the changes in the number of pre-crRNAs $\Delta[u]$ and crRNA $\Delta[p]$, upon CasE overexpression:

$$\Delta[p] = -\frac{\lambda_u}{\lambda_p}\Delta[u]$$

In the equation above, the minus sign indicates that the decrease in the number of unprocessed transcripts (pre-crRNAs) is accompanied by an increase in the number of processed transcripts (crRNAs).

From the above equation follows that crRNA increase is directly proportional to pre-crRNA decrease, with a large constant of proportionality that is equal to 100 ($\lambda_u/\lambda_p \sim 100$ - see the previous section). This large constant of proportionality explains the experimentally observed large amplification of crRNA upon CasE overexpression. That is, according to the equation, ~10 molecule decrease in pre-crRNA ($\Delta[u] \sim 10$), leads to two orders of magnitude larger increase in crRNA ($\Delta[p] \sim 1000$), as observed in the experiments. Therefore, the equation shows that the system acts as a strong linear amplifier, where the increase of crRNA is directly proportional to the decrease of pre-crRNA, and where a small number of pre-crRNAs is amplified to a large number of crRNAs. This large amplification directly contributes to efficient transition of the system from "OFF" state (with only few crRNA molecules) to "ON" state (with a large number of crRNA molecules).

4 Kinetics of crRNA generation

We next discuss kinetics of crRNA accumulation, in order to understand how fast the system can achieve crRNA levels that are sufficient for protection against foreign DNA invasion. One should note that the steady-state regime may not be directly relevant for system function under natural conditions, where the amount of generated crRNA immediately after induction (i.e., for example, after virus infection) may be more relevant. While it is hard to experimentally assess kinetics of the transcript accumulation, this analysis can be readily done through mathematical modeling, which will be discussed below.

We below consider what happens if transcription of both *cas* genes and CRISPR array is activated. This analysis is motivated by reported repression of *cas* and (to a smaller extent) CRISPR promoters by H-NS, and by a (widely accepted) model which proposes that the system is induced by abolishing this repression (see e.g. [5]). Activation of *cas* genes and CRISPR array transcription leads to increasing both pre-crRNA to crRNA processing rate and CRISPR transcription rate. One should note that the analysis discussed in the previous subsection corresponds to only increase of pre-crRNA to crRNA processing rate (as relevant for CasE overexpression experiments in which the transcript numbers were quantitated).

In Figure 3, we show the kinetics of crRNA accumulation for the parameters which are likely close to the natural system induction [11]. As can be seen, we have analyzed crRNA accumulation both deterministically (the blue curve) and stochastically (the magenta curves). We perform the stochastic simulations

since there are only few pre-crRNA and crRNA molecules before the system is induced, and since the number of pre-crRNA molecules can become even smaller after the system induction. However, it turns out that the stochastic and the deterministic results are in agreement with each other (see e.g. Fig. 3), which validates that the simple analytic expressions that we derive (see the previous section) can be used to describe the system.

One should note that experimental data indicate that repression by H-NS of the *cas* promoter is much stronger compared to the repression of the CRISPR array [5, 10]. It is therefore likely that activation of *cas* genes upon abolishment of H-NS repression is much larger than the activation of CRISPR array transcription, which is reflected by the choice of the induction parameters in Fig. 5. One should also note that the increase of crRNA steady-state values obtained in the figure is consistent with the values measured in experiments in which H-NS repression of *cas* and CRISPR promoters is abolished [5, 10]. That is, both Fig. 3 and the experiments in which H-NS repression is abolished show that the amount of crRNA increases for about two orders of magnitude. This provides another argument that the induction parameters used in Fig. 5 are likely close to the conditions relevant for natural system induction.

Fig. 3 shows that the steady state is reached relatively slowly, i.e. ~300 min after the system induction. However, when a virulent phage infects *E. coli*, the cell lysis is typically complete much before 300 min post-infection; e.g. for the well known *E. coli* T7 and T3 phages, the cell lysis starts at ~20 min post-infection, with complete shot-off of host functions occurring much earlier [12]. Therefore, crRNA values soon after the system induction (e.g. at ~20min post-induction) may be more relevant for the defense against foreign DNA than the steady state crRNA levels. In Fig. 5, we see that the transcript amounts at ~20 min. post-induction (~200 transcripts) are much higher compared to crRNA levels that were experimentally shown to provide a partial protection against bacteriophage infection (~10 crRNA transcripts as per [6]). Therefore, the results strongly suggest that activation of *cas* expression and CRISPR array leads to a rapid accumulation of crRNA, which can provide an effective protection against phage infection. Consequently, induction of CRISPR/Cas system also involves a rapid transition from unproductive to productive state of the system.

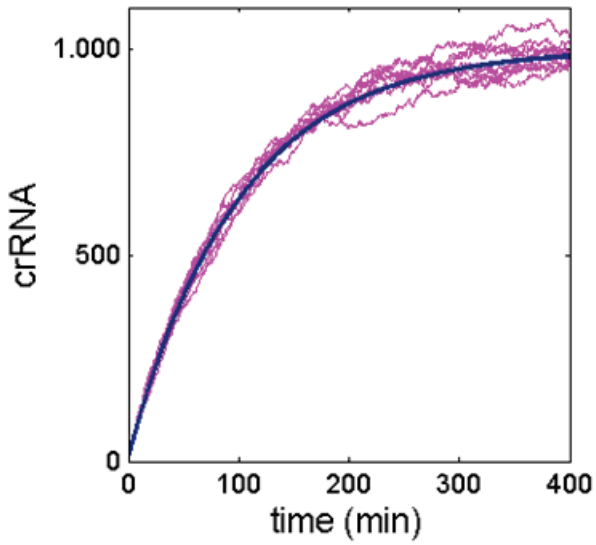


Figure 3. Kinetics of crRNA accumulation. The figure shows how crRNA changes as a function of time, when pre-crRNA to crRNA processing rate is increased for two orders of magnitude, while CRISPR transcription rate is increased two times. Horizontal axis corresponds to time post-induction, while the vertical axis corresponds to the number of crRNA transcripts.

5 Steady state stabilities

We above analyzed a sophisticated bacterial immune system (CRISPR/Cas), which uses a mechanism at the level of transcript processing to protect bacterial cell against virus infection. We have seen that, through this mechanism, the system can exhibit a fast transition from the "OFF" state (in which a cell is not protected against foreign DNA) to the "ON" state (in which there is a sufficient amount of toxic molecule to provide protection). In this subsection we briefly analyze if there are additional, more subtle, principles that determine the system design. We argue that increased stability of the steady state may be an example of such principle. Such increased stability of the steady state would prevent large fluctuations of the poison molecule (crRNAs in the case of CRISPR/Cas systems) that may be unmatched by the amount of the antidote.

An important control element of CRISPR/Cas system is fast non-specific degradation of pre-crRNA by an unidentified nuclease; as discussed above, this fast processing is a major element that allows fast transition of the system from "OFF" to "ON" state. In addition, it is straightforward to see that this fast non-

specific degradation increases stability of the steady state of the system: For example, if there is a perturbation which increases steady-state concentration of pre-crRNA, larger transcript decay will lead to a faster diminishing of this perturbation. Therefore, a major control element of CRISPR/Cas response also directly leads to increased stability of the steady state of the system.

6 Synthetic biology applications

Finally, we briefly analyze how our study of CRISPR transcript processing can be used in synthetic biology applications. For this, we go back to the experiments in which pre-crRNA to crRNA processing rate is increased. As discussed above, this increase results in a large product gain, where crRNA is the product. The large product gain is achieved despite small substrate amounts, which moreover further decrease as the system is induced. This provides a motivation for investigating whether one can make a synthetic system that is able to produce a large amount of product, from substrate that is consistently kept at low amounts; this would arise, for example, when we need to produce a lot of useful molecules from potentially toxic substrate.

With the goal of constructing such synthetic biology system, we now go back to the scheme of CRISPR transcript processing, where the substrate and the product are now any molecules that decay with some rates. In Figure 4, we show the generalized scheme for the product generation. According to the scheme, the substrate is generated with certain rate, and then converted to the product with some other rate. Our goal is finding an optimal induction strategy, so as to achieve the maximal product gain, while keeping the substrate amounts at low level.

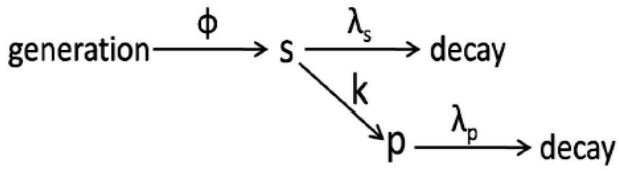


Figure 4. The generalized scheme for the product generation. The substrate s is generated with rate ϕ, and is consequently either degraded with rate rate λ_s, or is processed to product p with rate k; product p is then degraded with rate λ_p.

By using the scheme above, one can show [13] that the maximal product increase depends only on the increase of the substrate to product processing rate:

$$\left(\frac{p'}{p}\right)_{max} = \frac{k'}{k}$$

where the prime quantities correspond to the values after the system induction. However, to achieve this maximal increase, the substrate generation rate has to increase as well, according to the following equation:

$$\frac{\varphi'}{\varphi} \sim 1 + \frac{\frac{k'}{k} - 1}{\frac{\lambda_s}{\lambda_p} + 1}$$

Furthermore, from the equation we see that the required increase of the generation rate is inversely proportional to the ratio of the substrate and the product decay rates. We therefore obtain a surprising result that, within the general scheme analyzed above, the large substrate decay rate leads to a more efficient product generation [13].

Interestingly, a large ratio of the substrate to product decay rate, which follows from the system optimization, is exactly what happens in CRISPR/Cas system. In fact, if we use the decay rates measured for CRISPR/Cas, we obtain that the optimal increase of the processing rate is much larger compared to the optimal increase of the generation rate. This is consistent with a much stronger experimentally measured repression of Cas promoters compared to CRISPR promoters. It therefore appears that CRISPR/Cas system is optimized to produce large crRNA amounts, while pre-crRNA is kept low, which is an additional argument in favor of the design principles discussed above.

7 Conclusion

We have seen that the fast transition from "OFF" to "ON" state in CRISPR/Cas system is exhibited at the level of transcript processing, and crucially depends on fast non-specific degradation of pre-crRNA by an unidentified nuclease. Consequently, this nuclease is a major control element of CRISPR/Cas response. The large decay rate of pre-crRNA also increases stability of the steady-states for this system, which may be another important principle in the design of the system. Further study of the bacterial immune systems may lead to discovery of more such principles, which may be useful not only for understanding of endogenous systems, but also for construction of useful synthetic gene circuits. An example of this, which we analyzed here, is a possibility for generating a large amount of useful product, from small amounts of potentially toxic substrate. Interestingly, we also found that.

Acknowledgements

This work is supported by Marie Curie International Reintegration Grant within the 7th European community Framework Programme (PIRG08-GA-2010-276996) and by the Ministry of Education and Science of the Republic of Serbia under project number ON173052. We thank Konstantin Severinov for useful discussions on CRISPR/Cas systems.

References

1. Kobayashi, I., *Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution.* Nucleic Acids Res, 2001. **29**(18): p. 3742-56.
2. Makarova, K.S., et al., *A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.* Biol Direct, 2006. **1**: p. 7.
3. Barrangou, R., et al., *CRISPR provides acquired resistance against viruses in prokaryotes.* Science, 2007. **315**(5819): p. 1709-12.
4. Al-Attar, S., et al., *Clustered regularly interspaced short palindromic repeats (CRISPRs): the hallmark of an ingenious antiviral defense mechanism in prokaryotes.* Biol Chem, 2011. **392**(4): p. 277-89.
5. Pul, U., et al., *Identification and characterization of E. coli CRISPR-cas promoters and their silencing by H-NS.* Mol Microbiol, 2010. **75**(6): p. 1495-512.
6. Pougach, K., et al., *Transcription, processing and function of CRISPR cassettes in Escherichia coli.* Mol Microbiol, 2010. **77**(6): p. 1367-79.
7. Brouns, S.J., et al., *Small CRISPR RNAs guide antiviral defense in prokaryotes.* Science, 2008. **321**(5891): p. 960-4.
8. Diez-Villasenor, C., et al., *Diversity of CRISPR loci in Escherichia coli.* Microbiology, 2010. **156**(Pt 5): p. 1351-61.
9. Poranen, M.M., et al., *Global changes in cellular gene expression during bacteriophage PRD1 infection.* J Virol, 2006. **80**(16): p. 8081-8.
10. Westra, E.R., et al., *H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO.* Mol Microbiol, 2010. **77**(6): p. 1380-93.
11. Djordjevic, M., M. Djordjevic, and K. Severinov, *CRISPR transcript processing: a mechanism for generating a large number of small interfering RNAs.* Biol Direct, 2012. **7**: p. 24.
12. Kruger, D.H. and C. Schroeder, *Bacteriophage T3 and bacteriophage T7 virus-host cell interactions.* Microbiol Rev, 1981. **45**(1): p. 9-51.
13. Djordjevic, M. and M. Djordjevic, *A simple biosynthetic pathway for large product generation from small substrate amounts.* Phys Biol, 2012. **9**(5): p. 056004.

Improved Method for Transcription-start Site Prediction in Bacteria

Marko Djordjević [a]
Faculty of Biology, University of Belgrade, Studentski trg 16, Belgrade, Serbia

Magdalena Djordjević [b]
Institute of Physics Belgrade, University of Belgrade, Serbia

Abstract

Promoter prediction in bacteria is a classical bioinformatics problem, where available methods for regulatory element detection exhibit a very high number of false positives. To start addressing this problem, we systematically analyzed sigma 70 promoter elements in *E. coli*, where we used a Monte-Carlo based procedure (Gibbs Search) to de-novo align promoter elements for more than 300 experimentally detected sigma 70 transcription start sites. We significantly improved alignment of the promoter elements, as judged by correspondence with biophysical interaction data. We also focused on conserved sequences upstream of -10 element (so called -15 element), which were previously not included in

[a] e-mail address: dmarko@bio.bg.ac.rs
[b] e-mail address: magda@ipb.ac.rs

transcription start site searches. We used the aligned promoter elements to improve the information-theory method for promoter recognition, and showed that this improvement significantly (~ 50%) reduces the number of false positives, though their number is still quite high. We also showed that this improvement can correctly identify strong promoters in a newly sequenced bacteriophage genome.

1 Introduction

Bacterial RNA polymerase is a central enzyme in cell, and initiation of transcription by bacterial RNA polymerase is a major point in gene expression regulation. Core RNA polymerase cannot by itself initiate transcription, so a complex between RNA polymerase core and a σ factor, which is called RNA polymerase holoenzyme (RNAP) is formed. A major σ factor, which is responsible for transcription of housekeeping genes, is called σ^{70} in *E. coli* and σ^{A} in a number of other bacteria. In this work, we will concentrate on recognition of σ^{70} (σ^{A}) promoter elements [1]. Accurate recognition of transcription start sites is a necessary first step in understanding transcription regulation, which is in-turn necessary for understanding regulation of gene expression. Due to that, bioinformatic recognition of bacterial promoters is considered a major problem in bioinformatics. However, available methods for transcription start site detection show poor accuracy - in particular they lead to a very high number of false positives [2].

A necessary step in addressing this problem is achieving a quantitative understanding of promoter specificity, i.e. of regulatory elements that define bacterial promoter. However, aligning the promoter elements presents in-itself a highly non-trivial bioinformatic task due to complex structure of bacterial promoter. In particular, the main elements that determine promoter recognition are -35 element ($^{-35}$TTGACA^{-30}, where the coordinates in the superscript are relative to the transcription start site), -10 element ($^{-12}$TATAAT^{-7}), the spacer between these two elements, and the extended -10 element ($^{-15}$TG^{-14}) [3]. More recently, it was also noted that the entire region upstream of -10 element (spanning from coordinates -15 to -12) is important for transcription initiation, which is termed -15 element.

A major problem with the existing collections of the promoter elements is due to the following: *i*) they are based on initial alignments of a small collection of promoter elements which were performed `by eye' [4-7] *ii*) accurate aligning of -35 element is complicated by both variable distance from -35 element and by a lower conservation of this element [5] *iii*) it is non-trivial to produce an alignment with sufficient accuracy for analyzing -15 element, given a weaker conservation of this element compared to both -10 and -35 elements [7]. Our goal here is to perform a systematic `de-novo' alignment of the promoter elements on large collection of more than 300 experimentally confirmed σ^{70} transcription start sites in *E. coli*. We will then integrate the obtained alignments in an infor-

mation-theory framework for transcription start site prediction. The final goal is to assess to what extent such improved alignment contributes to the accuracy of σ^{70} promoter detection.

2 Promoter element alignment

To evade biases in alignment, we directly start from experimentally determined transcription start sites in genome [8]. Our strategy is to use a Gibbs search algorithm for unsupervised alignment of promoter elements, which we in the end improve through supervised search by weight matrices defined through the Gibbs algorithm. Our approach is to first align -10 element, and to consequently use this element as an anchor to align -35 element. Alignment of other relevant elements (spacer and -15 element) is directly determined once -10 element and -35 element are aligned. This approach is described in detail in [9].

To align -10 elements, we use the assembly of transcription start sites from RegulonDB database [8]. For our alignment we select only experimentally verified transcription start sites, i.e. we disregard all transcription start sites that are either not experimentally validated, or correspond to alternative σ factors. This selection results in the total of 342 σ^{70} transcription start sites, and we use the obtained start sites in order to extract DNA segments that correspond to positions -17 to -2, relative to the transcription start sites. These positions were chosen having in mind that the position of -10 element can deviate for 5bps, relative to its canonical position (-12 to -7) [10].

To identify the 6 bp long -10 elements within the selected DNA segments, we used the Gibbs sampler [11-12]. The algorithm allows us to perform an unsupervised search, i.e. we use no prior information on sequence specificity of -10 box. Some of the initial 342 segments were found not to contain a recognizable -10 box (possibly due to database miss-assignments); consequently, the search resulted in the identification of 322 aligned -10 boxes, which were used in further analysis.

To identify -35 elements, we started from the aligned -10 elements, and selected DNA segments that correspond to range from 16 to 25 bps from the upstream most base in the aligned -10 box. This range is based on the fact that -35 element is 6bp long and that the spacer length between -35 and -10 element is 15 to 19 bps. We again used the Gibbs sampler to search for 6bps long overrepresented motifs within these segments. The search resulted in a motif with the consensus sequence 'GTTGAC'; this motif is evidently shifted for 1bp relative to the established consensus of -35 element ('TTGACA'). This shift is not surprising given that *i*) The downstream-most base of -35 element shows relatively low

conservation, *ii*) there is a fairly good conservation of the base-pair immediately upstream of -35 element (see Table 1), *iii*) it is common that a Gibbs search results in motifs that are shifted relative to their optimal alignment [11].

We therefore manually shift the alignment obtained by Gibbs search for 1bp, so that it coincides with the established consensus, and construct a weight matrix for such realigned -35 motif. To insure that the optimal alignment is indeed selected, we identify the motif with the highest weight matrix score on each of the original segments. Those motifs then present our final alignment for -35 elements. Once we aligned -10 element and -35 element, it is straightforward to sample distribution of the spacer lengths. Similarly, once we aligned -10 element, -15 element spans from 3 bases upstream of -10 element to the upstream most base of -10 element.

3 Specificity of the aligned promoter elements

Specificities of the aligned promoter elements are shown in Figure 1, which is generated by EnoLogos [13].

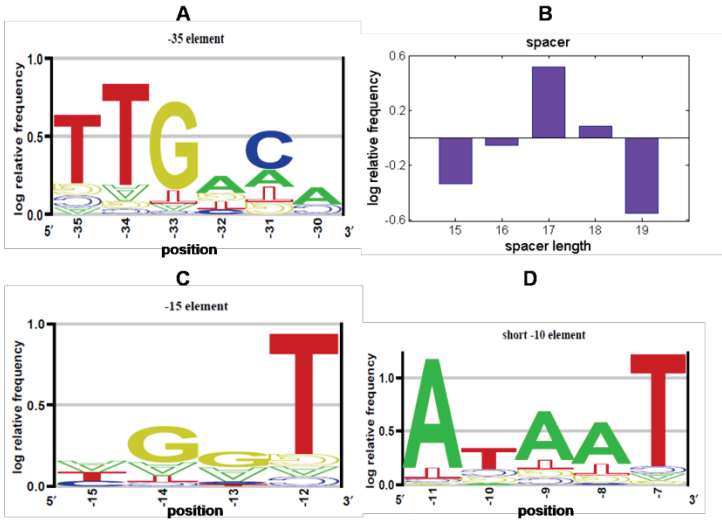


Fig1: Logarithm of the ratio of the base frequencies in the alignment, relative to the background base frequencies is shown in the figure. For spacers (Figure 1B), log ratios are also presented, where the background distribution is equiprobable. Sequence logos correspond to specificities of **A)** -35 element **B)** Spacer between -35 and -10 element **C)** -15 element **D)** short -10 element. Figure adopted from [9].

As described in [9], the overrepresentation of -35 element bases obtained from our alignment (Table 1 and Figure 1) is consistent with the available data on interactions between σ^{70} and -35 element [14]: The largest overrepresentation is obtained for bases -35, -34, -33 and -31, which are bound to σ subunit residues with hydrogen bonds; the overrepresentation is notably smaller for bases -32 and -30 which interact with σ^{70} with weaker van der Waals interactions. Finally, there is a statistically significant overrepresentation of G at position -36; this might seem unexpected, since position -36 is not part of -35 element; however, this conservation is consistent with the interaction data that indicate van der Waals interactions between -36 and σ^{70} residues [14].

We note that a recent alignment of -35 elements presented in [15] shows notable discrepancies with -35 element alignment presented here. Specifically, in [15] base 'C' at position -31 is significantly less conserved compared to 'A' at -32; this is inconsistent with the available data on interactions between σ^{70} and -35 element which indicate that base -31 interacts with σ^{70} through hydrogen bonds, while interactions with position -32 involve weaker van der Waals interactions. Furthermore, in [15] bases 'A' and 'T' show a larger conservation compared to 'C' and 'A' at positions -31 and -30, which is inconsistent with both the interaction data [14] and with -35 element consensus ('$^{-35}$TTGACA^{-30}') established through previous studies [3]; this is in contrast to our alignment where consensus '$^{-31}$C' and '$^{-30}$A' are clearly distinguished from the other bases at positions -31 and -30. Consequently, we expect that our alignment will lead to a substantially improved description of the promoter specificity.

The inferred specificity of -10 element is also consistent with available biophysical data: We see that the largest conservation corresponds to positions -11 and -7, which were shown in a number of studies to be of special importance for σ^{70}-ssDNA interactions (see e.g. [16-17]). On the other hand, mutations at position -10 showed no notable effect on σ^{70}-ssDNA binding [17], consistent with the smallest base overrepresentation at this position.

We next briefly address specificity of the binding positions within -15 motif. We first note a high degeneracy at position -15 where bases T and C are similarly overrepresented (1.18 and 1.14 relative to the background frequencies). Therefore, it is more appropriate to represent the extended -10 motif with a weight matrix, or qualitatively with a degenerate consensus, than with a consensus sequence. Next, we note a conservation of base 'G' at position -13, which appears at the frequency that is 1.4 times larger than the background frequency, which is statistically highly significant ($P \sim 10^{-3}$). We also note that conservation of the base at position -13 is larger than conservation at -15, which is a canonical base within the extended -10 motif (the 'T' in 'TG'). Conservation of base -13

at this position had not been reported before. Actually, the consensus sequence for the extended -10 motif is presented in the literature as 'TGn', where 'n' at position -13 indicates no conservation [3]. Consequently, we conclude that -15 element presents a conserved stretch of sequence, which has to be included in promoter search for a complete description of promoter specificity.

4 Transcription start site prediction

The aligned promoter elements are used as an input for the information-theory method for regulatory element detection. We start from a collection of aligned sequences. For each position in the alignment we determine frequency with which each of the four bases occurs. We also determine background base frequency by sampling frequency by which each of the four bases occurs in *E. coli* intergenic sequences; background frequencies were sampled from intergenic sequences, since transcription start sites are located within them. Weight matrix elements $w_{i,\alpha}$ that correspond to base present at position *i* in the motif are given by [18]:

$$w_{i,\alpha} = \log\left(\frac{n\nu_{i,\alpha} + p_\alpha}{p_\alpha\left(n+1\right)}\right),$$

where *n* is the total number of motifs in the alignment from which the weight matrix is inferred, $\nu_{i,\alpha}$ is number of times that base α appears at position *i* divided by *n*, and p_α is background frequency of base α. One should note that the addition of p_α in the numerator of the logarithm corresponds to so-called pseudo-counts, which become important for small datasets; note that for large *n* (as is the case for our dataset) the above expression approximately reduces to the log ratio of base frequency and background frequency.

A similar expression is used for weights corresponding to different spacer lengths: $w_i = \log\left(\nu_i / 0.2\right)$, where w_i is the weight corresponding to the spacer of length *i* ($i \in [15,..,19]$), while equiprobable background frequencies (0.2) were taken. Weight matrix score for the entire promoter than corresponds to sum of the weight matrix scores for all of the promoter elements (-35 element, -15 element, -10 element) and the weight corresponding to the spacer length. The actual search for the promoter elements in a given stretch of DNA sequence is done in the following way: for each position of -10 element, we determine which of the four spacer lengths corresponds to maximal weight matrix score

for -35 element. For such (maximal score) -35 element, we calculate the total promoter score for a given sequence.

The procedure described above allowed using our improved description of sigma 70 promoter specificity, within the information-theory method for promoter recognition. As described above, this improved specificity includes not only a description of -15 element - which was previously not included in promoter searches - but also a significantly improved specificity of -35 element. We next wanted to test how much improving the sequence specificity improves accuracy of transcription start site predictions. To that end, we compared our method with the standard procedure for transcription start prediction, as described in [6].

For the purpose of the comparison, we divided the initial set of ~320 experimentally determined σ70 transcription start sites in two groups: The first group of ~160 transcription start sites presents the training set; we consequently used the aligned promoter elements from this set to construct the weight matrices. The second group (of equal size) presents the testing set, which is used to estimate the number of false positives for both our method and the method in [6]. To estimate the number of false positives, we randomly permuted nucleotides in *E. coli* intergenic regions - note that this randomizes the sequence, while preserving the nucleotide content. The number of false positives is than estimated according to the number of hits in such randomized sequence.

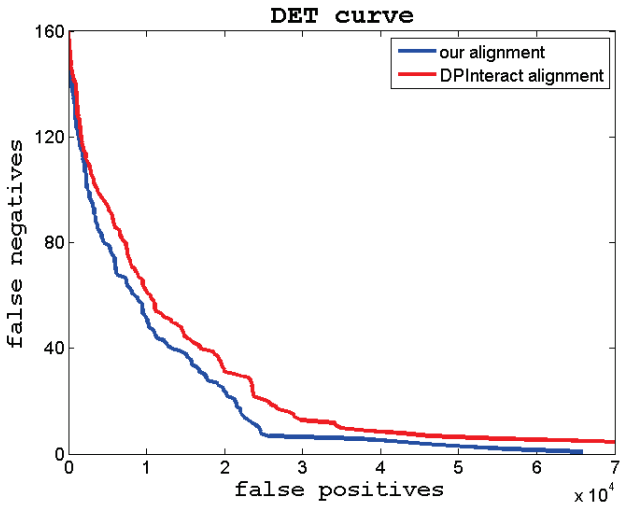


Figure 2: DET (Detection Error Tradeoff) curve, which presents comparison of the method based on our alignment, and the method in [6] that uses the alignment from DPInteract database. The vertical axis presents the number of false positives, which is based on the number of correctly classified sequences in the testing set. The horizontal axis presents the number of false negatives, which is estimated based on the number of hits in the randomized intergenic regions. The blue line and the red line correspond, respectively, to our procedure and the procedure from [6].

The results of the above comparison are shown in Figure 2, which presents DET (Detection Error Tradeoff) curve for comparison of the two methods. The number of false positives is shown on the vertical axis, while the estimate of the false positive number is shown on the horizontal axis. The two curves in Figure 2, are produced by varying the detection threshold. We see that across the entire range (for any detection threshold), our method (blue line) shows a significantly better false positive/false negative tradeoff compared to the standard method from [6] (red line). In particular, for the standard threshold choice, which leads to ~5% false positive value, our method leads to ~50% reduction in the number of false positives compared to the method from [6]. From Fig. 2, one can however see that even this reduction still leads to a large number of false positives. Reducing this large number of false positives will be a subject of our future work, where we hypothesize that further improvement in the accuracy requires explicitly taking into account kinetic effects in transcription initiation.

5 Conclusion

Accurate promoter prediction in bacteria is crucial not only as the first step in understanding transcription regulation, but also as an important ingredient in other bioinformatic applications such as gene and operon prediction. Despite being a classical bioinformatics problem, current methods for transcription start site prediction lead to very high number of false positives. We here investigated to what extent an improved description of promoter specificity can increase accuracy of promoter predictions. To that end, we used a Monte-Carlo based procedure to align a large number of experimentally determined promoters, without any prior bias. We showed that this procedure leads to a significantly improved specificity of -35 element, as judged by comparison with biophysical interaction data. Furthermore, we also quantified specificity for additional elements that determine promoter activity, in particular -15 element which is located upstream of -10 element. We next integrated this improved description of the promoter specificity in an information-theoretic framework for transcription start site detection. We showed that this leads to a significant (~50%) reduction in the number of false positives. The false positive number is however still high, which is likely a consequence of the complex physical nature of transcription initiation, and which will be explored in our future research.

Acknowledgements

This work is supported by Marie Curie International Reintegration Grant within the 7th European community Framework Programme (PIRG08-GA-2010-276996) and by the Ministry of Education and Science of the Republic of Serbia under project number ON173052.

References

1. Borukhov, S., Nudler, E.: RNA polymerase holoenzyme: structure, function and biological implications. Curr Opin Microbiol 6, 93-100 (2003)
2. Stormo, G.D.: DNA binding sites: representation and discovery. Bioinformatics 16, 16-23 (2000)
3. Hook-Barnard, I.G., Hinton, D.M.: Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. Gene Regulation and Systems Biology 1, 275 (2007)
4. Wang, H., Benham, C.J.: Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. BMC bioinformatics 7, 248 (2006)
5. Huerta, A.M., Collado-Vides, J.: Sigma 70 Promoters in Escherichia coli: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. J Mol Biol 333, 261-278 (2003)
6. Robison, K., McGuire, A., Church, G.: A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. Journal of molecular biology 284, 241-254 (1998)
7. Mitchell, J.E., Zheng, D., Busby, S.J.W., Minchin, S.D.: Identification and analysis of 'extended–10'promoters in Escherichia coli. Nucleic acids research 31, 4689 (2003)
8. Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñiz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J.S., López-Fuentes, A.: RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). Nucleic acids research 39, D98 (2011)
9. Djordjevic, M.: Redefining Escherichia coli sigma(70) promoter elements: -15 motif as a complement of the -10 motif. Journal of bacteriology 193, 6305-6314 (2011)
10. Harley, C.B., Reynolds, R.P.: Analysis of E. coli promoter sequences. Nucleic Acids Res 15, 2343-2361 (1987)
11. Lawrence, C., Altschul, S.: Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262, 208 (1993)
12. Thompson, W., Rouchka, E.C., Lawrence, C.E.: Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res 31, 3580-3585 (2003)
13. Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D., Benos, P.V.: enoLOGOS: a versatile web tool for energy normalized sequence logos. Nucleic Acids Res 33, W389-392 (2005)
14. Campbell, E.A., Muzzin, O., Chlenov, M., Sun, J.L., Olson, C.A., Weinman, O., Trester-Zedlitz, M.L., Darst, S.A.: Structure of the bacterial RNA polymerase promoter specificity [sigma] subunit. Molecular cell 9, 527-539 (2002)
15. Shultzaberger, R.K., Chen, Z., Lewis, K.A., Schneider, T.D.: Anatomy of Escherichia coli s 70 promoters. Nucleic Acids Res 35, 771-788 (2007)
16. Matlock, D.L., Heyduk, T.: Sequence determinants for the recognition of the fork junction DNA containing the-10 region of promoter DNA by E. coli RNA polymerase. Biochemistry 39, 12274-12283 (2000)
17. Fenton, M.S., Gralla, J.D.: Function of the bacterial TATAAT-10 element as single-stranded DNA during RNA polymerase isomerization. Proceedings of the National Academy of Sciences of the United States of America 98, 9020-9025 (2001)
18. Hertz, G.Z., Stormo, G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15, 563-577 (1999)

On Ultrametricity in Bioinformation Systems

Branko Dragovich [a]
Institute of Physics, University of Belgrade, Belgrade, Serbia

Abstract

Ultrametric aspects of bioinformation systems are considered. In particular, p-adic ultrametricity of the genetic code was investigated. We successfully applied p-adic distance to the space of 64 codons and 20 amino acids. We propose to use p-adic distance to investigate similarity (nearness) between strings of DNA, RNA, amino acids and other bioinformation systems. We present here a brief review of performed research and indicate some prosperous future investigations.

1 Introduction

In this paper, we want to emphasize the role of ultrametric distance, and in particular, p-adic one. Namely, some parts of a biological system can be considered simultaneously with respect to different metrics – the usual Euclidean metric, which measures spatial distances, and ultrametrics, which measure nearness related to some bioinformation properties.

The general notion of metric space (M, d) is introduced in 1906 by Maurice Fréchet (1878–1973), where M is a set and d is a distance function. Distance d is a real-valued function of any two elements $x, y \in M$

[a] e-mail address: dragovich@ipb.ac.rs

which must satisfy the following three properties: (i) $d(x,y) = 0 \Leftrightarrow x = y$, (ii) $d(x,y) = d(y,x)$, $(iii) d(x,y) \leq d(x,z) + d(z,y)$, where last property is called triangle inequality. An ultrametric space is a metric space which satisfies strong triangle inequality, i.e.

$$d(x,y) \leq \max\{d(x,z), d(z,y)\}. \tag{1}$$

Word *ultrametric* is introduced in 1944 by Marc Krasner (1912–1985), although examples of ultrametric spaces have been known earlier under different names. An important class of ultrametric spaces contains fields of p-adic numbers, which are introduced in 1897 by Kurt Hensel (1861–1941). Taxonomy, which started 1735 by Carl Linné (1707–1778) as biological classification with hierarchical structure, is another significant example of ultrametricity.

Here we mainly consider some aspects of the genetic code using p-adic ultrametric distance. The genetic code (GC) is connection between 64 codons, which are building blocks of the genes, and 20 amino acids, which are building blocks of the proteins. In addition to coding amino acids, a few codons code stop signal, which is at the end of genes and terminates process of protein synthesis. Codons are ordered triples composed of C, A, U (T) and G nucleotides. Each codon presents an information which is related to the use of one of the 20 standard amino acids or stop signal in synthesis of proteins. It is obvious that there are $4 \times 4 \times 4 = 64$ codons. For molecular biology and the genetic code, one can refer to [1].

From mathematical point of view, the GC is a mapping of a set of 64 elements onto a set of 21 element. There is in principle a huge number of possible mappings (about 10^{84}), but the genetic code is one definite mapping with a few (about 30) slight modifications. Hence, for modeling of the GC, the main problem is to find the corresponding structure of the space of 64 and 20 (or 21) elements. It will be demonstrated here that the set of 64 codons, and 20 amino acids, has p-adic structure, where $p = 5$ and $p = 2$. Some more detail expositions of p-adic approach to the genetic code are presented in [2, 3, 4, 5, 6] (see also [7] for a similar consideration). As review articles, we refer readers to [8] for application of ultrametricity and to [9, 10] for applications of p-adicity in physics and related topics.

Taking into account p-adic distance between constituents of two strings, we also consider modified Hamming distance.

2 p-Adic structure of the genetic code

p-Adic approach is based on the following idea. Codons, which code the same amino acid, should be treated close (near) each other in the information space. To quantify this closeness (nearness) we should use some distance. Ordinary distance is unappropriate. From insight to the table of the genetic code (see, e.g. Table 1) one can conclude that distribution of codons is like an ultrametric tree and it suggests use of p-adic distance.

Let us now introduce the following subset of natural numbers with respect to the base 5:

$$\mathcal{C}_5\,[64] = \{n_0 + n_1\,5 + n_2\,5^2\ :\ n_i = 1, 2, 3, 4\}\,, \tag{2}$$

where n_i are the corresponding digits different from zero. This is a three-digit expansion to the base 5, which is a prime number. The set $\mathcal{C}_5[64]$ contains 64 natural numbers. It is convenient to denote elements of $\mathcal{C}_5[64]$ by their digits to the base 5 in the following way: $n_0 + n_1\,5 + n_2\,5^2 \equiv n_0\,n_1\,n_2$. Here ordering of digits follows the expansion and it is opposite to the usual one.

We are now interested in 5-adic distances between elements of $\mathcal{C}_5\,[64]$. It is worth recalling p-adic norm between integers, which is related to the divisibility of integers by prime numbers. p-Adic distance between two integers can be understood as a measure of divisibility of their difference by p (the more divisible, the shorter distance). By definition, p-adic norm of an integer $m \in \mathbb{Z}$, is $|m|_p = p^{-k}$, where $k \in \mathbb{N}\bigcup\{0\}$ is degree of divisibility of m by prime p (i.e. $m = p^k\,m'$, $p \nmid m'$) and $|0|_p = 0$. This norm is a mapping from $\mathbb{Z}$ into non-negative rational numbers. One can easily conclude that $0 \le |m|_p \le 1$ for any $m \in \mathbb{Z}$ and any prime p.

p-Adic distance between two integers x and y is

$$d_p(x\,, y) = |x - y|_p\,. \tag{3}$$

Since p-adic norm is ultrametric, the p-adic distance (3) is also ultrametric, i.e. it satisfies inequality

$$d_p(x\,, y) \le\ \max\{d_p(x\,, z)\,, d_p(z\,, y)\}, \tag{4}$$

where x, y and z are any three integers.

5-Adic distance between two numbers $a, b \in \mathcal{C}_5\,[64]$ is

$$d_5(a,\,b) = |a_0 + a_1\,5 + a_2\,5^2 - b_0 - b_1\,5 - b_2\,5^2|_5\,, \tag{5}$$

where a_i, $b_i \in \{1, 2, 3, 4\}$. When $a \ne b$ then $d_5(a,\,b)$ may have three different values:

Table 1: The p-adic vertebrate mitochondrial genetic code.

111	CCC	Pro	211	ACC	Thr	311	UCC	Ser	411	GCC	Ala
112	CCA	Pro	212	ACA	Thr	312	UCA	Ser	412	GCA	Ala
113	CCU	Pro	213	ACU	Thr	313	UCU	Ser	413	GCU	Ala
114	CCG	Pro	214	ACG	Thr	314	UCG	Ser	414	GCG	Ala
121	CAC	His	221	AAC	Asn	321	UAC	Tyr	421	GAC	Asp
122	CAA	Gln	222	AAA	Lys	322	UAA	Ter	422	GAA	Glu
123	CAU	His	223	AAU	Asn	323	UAU	Tyr	423	GAU	Asp
124	CAG	Gln	224	AAG	Lys	324	UAG	Ter	424	GAG	Glu
131	CUC	Leu	231	AUC	Ile	331	UUC	Phe	431	GUC	Val
132	CUA	Leu	232	AUA	Met	332	UUA	Leu	432	GUA	Val
133	CUU	Leu	233	AUU	Ile	333	UUU	Phe	433	GUU	Val
134	CUG	Leu	234	AUG	Met	334	UUG	Leu	434	GUG	Val
141	CGC	Arg	241	AGC	Ser	341	UGC	Cys	441	GGC	Gly
142	CGA	Arg	242	AGA	Ter	342	UGA	Trp	442	GGA	Gly
143	CGU	Arg	243	AGU	Ser	343	UGU	Cys	443	GGU	Gly
144	CGG	Arg	244	AGG	Ter	344	UGG	Trp	444	GGG	Gly

- $d_5(a,\, b) = 1$ if $a_0 \neq b_0$,
- $d_5(a,\, b) = 1/5$ if $a_0 = b_0$ and $a_1 \neq b_1$,
- $d_5(a,\, b) = 1/5^2$ if $a_0 = b_0\,,\ a_1 = b_1$ and $a_2 \neq b_2$.

We see that the largest 5-adic distance between the above numbers is 1 and it is maximum p-adic distance on $\mathbb{Z}$. The smallest 5-adic distance on the space $\mathcal{C}_5\,[64]$ is 5^{-2}. Note that 5-adic distance depends only on the first two digits of different numbers $a, b \in \mathcal{C}_5\,[64]$.

Ultrametric space $\mathcal{C}_5[64]$ can be viewed as 16 quadruplets with respect to the smallest 5-adic distance, i.e. quadruplets contain 4 elements and 5-adic distance between any two elements within quadruplet is $\frac{1}{25}$. In other words, within each quadruplet, elements have the first two digits equal and third digits are different.

With respect to 2-adic distance, the above quadruplets may be viewed as composed of two doublets: $a = a_0\, a_1\, 1$ and $b = a_0\, a_1\, 3$ make the first doublet, and $c = a_0\, a_1\, 2$ and $d = a_0\, a_1\, 4$ form the second one. 2-Adic distance between codons within each of these doublets is $\frac{1}{2}$, i.e.

$$d_2(a,\, b) = |(3-1)\, 5^2|_2 = \frac{1}{2}\,, \qquad d_2(c,\, d) = |(4-2)\, 5^2|_2 = \frac{1}{2}\,. \tag{6}$$

By this way ultrametric space $\mathcal{C}_5[64]$ of 64 elements is arranged into 32 doublets.

Identifying nucleotides with digits in $\mathcal{C}_5[64]$ in the following way: C (cytosine) = 1, A (adenine) = 2, T (thymine) = U (uracil) = 3, G (guanine) = 4, we find one-to-one correspondence between codons in three-letter notation and three-digit $n_0\, n_1\, n_2$ number representation. Looking at Table 1 for the vertebrate mitochondrial genetic code one can easily see that 5-adic with 2-adic distances generate 32 doublets which are attached to 20 amino acids and one stop signal, i.e. now 32 elements are mapped onto 21 element. Note that nearness inside purines and pyrimidines, as well as between them, is described by 2-adic distance. Namely, 2-adic distance between pyrimidines C and U is $d_2(1, 3) = |3-1|_2 = 1/2$ and the distance between purines A and G is $d_2(2, 4) = |4-2|_2 = 1/2$. However 2-adic distance between C and A or G as well as distance between U and A or G is 1 (i.e. maximum).

By the above application of 5-adic and 2-adic distances to $\mathcal{C}_5\,[64]$ codon space we have obtained internal structure of the codon space in the form of doublets. Just this p-adic structure of codon space with doublets corresponds to the vertebrate mitochondrial genetic code, presented at Table 1.

Table 2: 20 standard amino acids with assigned 5-adic numbers.

11 Proline	21 Threonine	31 Serine	41 Alanine
12 Histidine	22 Asparagine	32 Tyrosine	42 Aspartate
13 Leucine	23 Isoleucine	33 Phenynalanine	43 Valine
14 Arginine	24 Lysine	34 Cysteine	44 Glycine
1 Glutamine	2 Methionine	3 Tryptophan	4 Glutamate

The other (about 30) known versions of the genetic code in living systems can be viewed as slight modifications of this mitochondrial code, which seems to be basic trinucleotide code.

At Table 2 we assigned numbers $x_0x_1 \equiv x_0 + x_1\,5$ to 16 amino acids which are assumed to be present in dinucleotide coding epoch, and $x_0 = 1, 2, 3, 4$ is attached to four late amino acids which were added during trinucleotide coding. Having these 20 5-adic numbers for amino acids one can consider distance between them and codons: there are 23 codon doublets which are at $\frac{1}{25}$ 5-adic distance with the corresponding 15 amino acids, i.e. codons within these doublets and related amino acids are at the same 5-adic distance. The other 5 a.a. are at $\frac{1}{5}$ distance with respect to their codon doublets.

Comparing Table 2 with temporal appearance of the 20 standard amino acids (see [11] and [4]) one can conclude that dinucleotide code preceded the trinucleotide one, which appeared adding third nucleotide. Then it becomes natural that the second nucleotide expresses chemical similarity between amino acids. Hence to characterize chemical nearness between amino acids it should be interchanged the first and second digit and then to take 5-adic distance between them.

3 p-Adically modified Hamming distance

Let $a = a_1\,a_2 \cdots a_n$ and $b = b_1\,b_2 \cdots b_n$ be two strings of equal length. Hamming distance between these two strings is $d_H(a, b) = \sum_{i=1}^{n} d(a_i, b_i)$, where $d(a_i, b_i) = 0$ if $a_i = b_i$, and $d(a_i, b_i) = 1$ if $a_i \neq b_i$. We introduce p-adically modified Hamming distance in the following way: $d_{pH}(a, b) =$

$\sum_{i=1}^{n} d_p(a_i, b_i)$, where $d_p(a_i, b_i) = |a_i - b_i|_p$ is p-adic distance between numbers a_i and b_i. When $a_i, b_i \in \mathbb{N}$ then $d_p(a_i, b_i) \leq 1$. If also $a_i - b_i \neq 0$ is divisible by p then $d_p(a_i, b_i) < 1$. In the case of strings as parts of DNA, RNA and proteins, this modified distance is finer and should be more appropriate than Hamming distance itself. For example, elements a_i and b_i can be nucleotides, codons and amino acids with above assigned natural numbers, and primes $p = 2$ and $p = 5$.

4 Concluding remarks

It is worth noting that codons can be viewed as three letter words of the four letter alphabet, and that proteins are some words of 20 letter alphabet. Then there is some analogy between information similarities of words in human language and natural codon language. Namely, in these both languages the first letters play more important role in the words meaning than the others. So the words with the same few letters have the same or very similar meaning. It follows that words of human language also have some ultrametricity.

There are many examples of hierarchical organization in living systems: from biomolecules to social networks, including hierarchical organization of brain functional network [12]. Natural way to investigate hierarchical systems is by ultrametrics, and particularly by p-adics.

Note that in our ultrametric approach to codon space there are no codon dynamics based on arithmetic operations, e.g. there is no summation of two codons which gives the third one. However, any codon can map itself to any other one by changes of its nucleotides, i. e. by changing its digits. Perhaps similar situation is with information processing in the brain. If so, then the brain is not like a computer which uses arithmetics, but a bioinformation system which works in a geometric way.

At the end, we can conclude that ultrametric approach to analyze bioinformation is at its beginning and has a prosperous future.

Acknowledgements

This work was partially supported by the Ministry of Education, Science and Technological Developments of the Republic of Serbia, contracts 173052 and 174012.

References

[1] J. D. Watson, T. A. Baker, S. P. Bell, A. Gann, M. Levine and R. Losick, *Molecular Biology of the Gene* (CSHL Press, Benjamin Cummings, San Francisco, 2004).

[2] B. Dragovich and A. Dragovich, "A p-adic model of DNA sequence and genetic code", p-Adic Numbers, Ultrametric Analysis and Applications **1** (1), 34–41 (2009); [arXiv:q-bio.GN/0607018v1].

[3] B. Dragovich and A. Dragovich, "p-Adic degeneracy of the genetic code", SFIN XX A1, 179–188 (2007); [arXiv:0707.0764 [q-bio.GN]].

[4] B. Dragovich and A. Dragovich, "p-Adic modelling of the genome and the genetic code", The Computer Journal **53** (4), 432–442 (2010); [arXiv:0707.3043v1 [q-bio.OT]].

[5] B. Dragovich, "Genetic code and number theory", [arXiv:0911.4014 [q-bio.OT]], (2009).

[6] B. Dragovich, "p-Adic structure of the genetic code", NeuroQuantology **9** (4), 716–727 (2011); [arXiv:1202.2353v1 [q-bio.OT]].

[7] A. Khrennikov and S. Kozyrev, "Genetic code on a diadic plane," Physica A: Stat. Mech. Appl. **381**, 265–272 (2007); [arXiv:q-bio/0701007].

[8] R. Rammal, G. Toulouse and M. A. Virasoro, "Ultrametricity for physicists", Rev. Mod. Phys. **58**, 765–788 (1986).

[9] L. Brekke and P. G. O. Freund, "p-Adic numbers in physics", Phys. Rept. **233**, 1–66 (1993).

[10] B. Dragovich, A. Yu. Khrennikov, S. V. Kozyrev and I. V. Volovich, "On p-adic mathematical physics", p-Adic Numbers, Ultrametric Analysis and Applications **1**, 1–17 (2009); [arXiv:0904.4205v1 [math-ph]].

[11] E. N. Trifonov, "The triplet code from first principles", J. Biomol. Struct. Dyn. **22**, 1–11 (2004).

[12] Zhao Zhou, Shi-Min Cai, Zhong-Qian Fu and Jie Zhang, "Hierarchical organization of brain functional network during visual task", Phys. Rev. E 84, 031923 (2011); [arXiv:1101.4773 [physics.bio-ph]].

Can We Use Standard Tools to Predict Functional Effects of Missense Gene Variations Outside Conserved Domains? TET2 Example

Branislava Gemović [a]
Vladimir Perović [b]
Sanja Glišić [c]
Nevena Veljković [d]
Centre for Multidisciplinary Research and Engineering,
Vinča Institute of Nuclear Sciences, University of Belgrade, Belgrade, Serbia

Abstract

The most common genetic variations in humans are Single Nucleotide Polymorphisms (SNPs), so predicting their associations with cancers is a significant issue. Here, we were particularly interested in SNPs occurring outside protein Conserved Domains

[a] e-mail address: gemovic@vinca.rs
[b] e-mail address: vladaper@vinca.rs
[c] e-mail address: sanja@vinca.rs
[d] e-mail address: nevenav@vinca.rs

(CDs) of TET2, a recently discovered epigenetic regulator involved in leukemogenesis. Functional effects of TET2 gene variations were assessed with four publicly available tools: PhD-SNP, MutPred, PolyPhen-2 and SIFT. The methods were tested on the dataset of 166 SNPs and somatic TET2 mutations, and separately on the subset of 69 variations outside TET2 CDs. Abilities of tested tools to separate neutral SNPs from pathogenic mutations were similar to previously reported on complete TET2 dataset. However, we observed significantly lower accuracy of predictions outside CDs, ranging from 0.54 to 0.62. Also, areas under the ROC curves were low, 0.51-0.55. Correlations between predictions and positions of variations inside/outside CDs were significant and high, 0.46-0.78. Low efficiency of commonly used tools in predicting functional effects of variations outside CDs emphasize the need for new or modified algorithms.

1 Introduction

The most frequent human genetic variations are SNPs, of which an important subset contains SNPs resulting in the amino acid substitutions (AAS). These mutations play one of the most important roles in cancer transformation [1, 2]. A number of tools have been developed to computationally predict which AAS have relevant phenotypic effect [for review see 3]. In this study we evaluated four widely used tools PhD-SNP [4], MutPred [5], PolyPhen-2 [6] and SIFT [7]. The stated tools use different protein features for predicting pathogenic effects of AAS. SIFT uses only evolutionary information, PhD-SNP combines it with sequence properties, while PolyPhen-2 and MutPred use a number of structural and functional data, in addition.

Several previous studies showed that more than 50% of cancer-associated mutations are positioned outside CDs [8, 9]. Also, extensive analysis of mutations in the important cancer-associated protein family, protein kinases, showed that numerous driver mutations are not in the kinase domains [10]. Nonetheless, performance evaluation of prediction tools has never been specifically focused on the effects of variations outside protein CDs.

TET2 is epigenetic regulator acting as an enzyme, normally converting 5-methylcytosine to 5-hydroxymethylcytosine in DNA [11]. It has been frequently mutated in all types of myeloid malignancies [12]. TET2 mutations predispose hematopoietic stem cells towards uncontrolled self-renewal and consequently development of myeloid malignancies [13, 14]. Even more,

mutations in TET2 are prognostic markers in acute myeloid leukemia [15] and play a role in leukemia transformation [16]. Having two well defined CDs and numerous AAS identified along entire sequence, TET2 represents a good candidate gene for pilot testing on the ability of published computational tools to discriminate between neutral SNPs and pathogenic mutations outside CDs.

2 Materials and Methods

Missense variations in TET2 gene were collected from literature, COSMIC [17] and dbSNP database [18].To label an AAS as a mutation, besides its association with a myeloid malignancy, we looked in original papers for evidence of its somatic nature. There were two criteria to label an AAS as a SNP: first included evidence in original papers of its presence in germline and the second implied described frequency of the polymorphism in healthy population. All-TET2 dataset contained 166 TET2 variations, of which 121 were mutations associated with myeloid malignancies. Also, we constructed a sub-dataset nCD-TET2 from all-TET2 that contained 69 variations outside TET2 CDs, 42 neutral SNPs and 27 mutations. TET2 CDs and non CD regions were determined from the relevant literature [19].

The pathogenicity of TET2 variations were predicted by the tools PhD-SNP [4], MutPred [5], PolyPhen-2 [6] and SIFT [7]. For all tools, we applied default parameters. Contrary to other three tools, PhD-SNP does not give probability scores as a result, so all statistical analyses for this method was done solely on the predictions. PolyPhen-2 and MutPred provide probability scores for a hypothesis that a variation is a damaging mutation and score of 0.5 was used as a predictions threshold. In the case of SIFT, variation is predicted to be a damaging mutation if the probability score is less than 0.05.

The performance of the four tools was assessed by three parameters: accuracy, sensitivity and specificity. For the additional evaluation of prediction tools, we constructed receiver operating characteristic (ROC) curves for both probability scores, where applicable, and predictions. The parameter used was area under the curve (AUC). Correlations between the predictions of tools and position of the variations inside/outside TET2 CDs were calculated using Spearman's rank correlation coefficients. For the determination of the significance of the results, we used chi-square test. The p-values were estimated in a two-tailed fashion. The significance threshold was p-value ≤ 0.01.

3 Results and Discussion

First, we evaluated the performance of PhD-SNP, MutPred, PolyPhen-2 and SIFT in predicting the pathogenicity of missense variants positioned outside TET2 CDs (Table 1). Although accuracies of PhD-SNP and MutPred were somewhat higher than accuracies of PolyPhen-2 and SIFT, the sensitivity and specificity of these tools were quite

	nCD-TET2 dataset			all-TET2 dataset		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
PhD-SNP	0.61	0.04	0.98	0.75	0.67	0.98
MutPred	0.62	0.04	1.00	0.52	0.34	1.00
PolyPhen-2	0.54	0.37	0.64	0.78	0.83	0.62
SIFT	0.55	0.52	0.57	0.78	0.85	0.58

unbalanced. So, we used AUC as additional measure of the performance of these four tools (Fig.1A). PhD-SNP, PolyPhen-2 and SIFT had extremely low AUC values ranging from 0.55 to 0.59 for probability scores and 0.51-0.55 for predictions. MutPred showed high discrepancy between AUC values of its probability scores (AUC=0.68) and predictions (AUC=0.52). This implies that predictions threshold of 0.5, suggested by authors, doesn't represent the optimal value for this particular dataset. But, although higher than for other three tools, performance of MutPred, still, cannot be considered satisfactory.

The accuracy of tested tools predicting pathogenicity of myeloid malignancies-associated variations positioned outside TET2 CDs was shown to be much lower than in the case of more comprehensive datasets, containing mutations not restricted to nCD-regions and originating from various diseases [20, 21]. So, we tested if our findings are specific for the TET2 variations, by evaluating the same tools on the complete all-TET2 dataset. As can be observed from Table 1, PhD-SNP, PolyPhen-2 and SIFT performances were in accordance with previously mentioned studies. Of note, MutPred prediction capacity on the all-TET2 dataset was significantly lower than reported by Thusberg et al. [20] and Li et al. [5]. We are speculating that this is, again, on the account of the predefined prediction threshold which is not appropriate, similarly to the nCD-TET2 dataset. Nevertheless, differences in AUC values for all tested tools between all-TET2 and nCD-TET2 datasets (Fig.1B), also, reflect decrease of their performance when dataset contains only variations outside CDs.

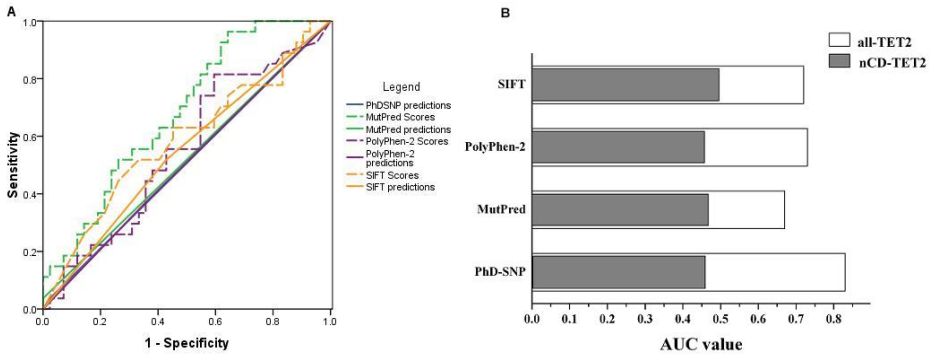


Figure 1: ROC analysis of PhD-SNP, MutPred, PolyPhen-2 and SIFT predictions of pathogenicity of nCD-TET2 variations. **A** ROC curves for probability scores and predictions (only predictions were available for PhD-SNP); **B** Difference between AUC values of predictions on all-TET2 and nCD-TET2 datasets

All tested tools base their predictions on the conservation of the amino acid position in a sequence, so we assumed that their predictions correlate significantly with the position of AAS in TET2 sequence, i.e. whether it is placed in the CD or not. To test this, we compared, pairwise, predictions of each tool and positions of variations in the CD/nCD (Table 2) and observed significant correlations ($p < 0.001$).

	PhD-SNP	MutPred	PolyPhen-2	SIFT
CD/nCD	0.78	0.46	0.65	0.52

Together, our results suggest that tested tools tend to use information about the position of variation in the protein CDs to annotate this variation as a mutation. On TET2 example, this is reflected by the accuracy of 0.95 of PolyPhen-2 and SIFT when we tested variations placed inside CDs (data not shown). But, tendency of these tools to annotate variations outside CDs as neutral SNPs can result in high number of false negatives and this can be the reason for the poor performance on our nCD-TET2 dataset.

In this pilot study, we intended to emphasize the importance of considering the information other than evolutionary in computational tools that predict disease related mutations in complex diseases.

Acknowledgements

This work is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant No. 173001).

References

[1] T.J. Ley, E.R. Mardis, L. Ding, B. Fulton, M.D. McLellan, K. Chen, D. Dooling, B.H. Dunford-Shore et al., *Nature* **456** (2008) 66.

[2] E.D. Pleasance, R.K. Cheetham, P.J. Stephens, D.J. McBride, S.J. Humphray, C.D. Greenman, I. Varela, M.L. Lin et al., *Nature* **463** (2010) 191.

[3] D.M. Jordan, V.E. Ramensky and S.R. Sunyaev, *Curr. Opin. Struct. Biol.* **20** (2010) 342.

[4] E. Capriotti, R. Calabrese and R. Casadio, *Bioinformatics* **22** (2006) 2729.

[5] B. Li, V.G. Krishnan, M.E. Mort, F. Xin, K.K. Kamati, D.N. Cooper, S.D. Mooney and P. Radivojac, *Bioinformatics* **25** (2009) 2744.

[6] I.A. Adzhubei, S. Schmidt, L. Peshkin, V.E. Ramensky, A. Gerasimova, P. Bork, A.S. Kondrashov and S.R. Sunyaev, *Nat. Methods.* **7** (2010) 248.

[7] P.C. Ng and S. Henikoff, *Nucleic Acids Res.* **31** (2003) 3812.

[8] P. Yue, W.F. Forrest, J.S. Kaminker, S. Lohr, Z. Zhang and G. Cavet, *Hum. Mutat.* **31** (2010) 264.

[9] T.A. Peterson, N.L. Nehrt, D. Park and M.G. Kann, *J. Am. Med. Inform. Assoc.* **19** (2012) 275.

[10] C. Greenman, P. Stephens, R. Smith, G.L. Dalgliesh, C. Hunter, G. Bignell, H. Davies, J. Teague et al., *Nature* **446** (2007) 153.

[11] M. Tahiliani, K.P. Koh, Y. Shen, W.A. Pastor, H. Bandukwala, Y. Brudno, S. Agarwal, L.M. Iyer et al., *Science* **324** (2009) 930.

[12] F. Delhommeau, S. Dupont, V. Della Valle, C. James, S. Trannoy, A. Massé, O. Kosmider, J.P. Le Couedic et al., *N. Engl. J. Med.* **360** (2009) 2289.

[13] K. Moran-Crusio, L. Reavie, A. Shih, O. Abdel-Wahab, D. Ndiaye-Lobry, C. Lobry, M.E. Figueroa, A. Vasanthakumar et al., *Cancer Cell* **20** (2011) 11.

[14] C. Quivoron, L. Couronné, V. Della Valle, C.K. Lopez, I. Plo, O. Wagner-Ballon, M. Do Cruzeiro, F. Delhommeau et al., *Cancer Cell* **20** (2011) 25.

[15] K.H. Metzeler, K. Maharry, M.D. Radmacher, K. Mrózek, D. Margeson, H. Becker et al., *J. Clin. Oncol.* **29** (2011) 1373.

[16] O. Abdel-Wahab, T. Manshouri, J. Patel, K. Harris, J. Yao, C. Hedvat, A. Heguy, C. Bueso-Ramos et al., *Cancer Res.* **70** (2010) 447.

[17] S.A. Forbes, N. Bindal, S. Bamford, C. Cole, C.Y. Kok, D. Beare, M. Jia, R. Shepherd et al., *Nucleic Acids Res.* **39** (2011) D945.

[18] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski and K. Sirotkin, *Nucleic Acids Res.* **29** (2001) 308.

[19] S.M. Langemeijer, R.P. Kuiper, M. Berends, R. Knops, M.G. Aslanyan, M. Massop, E. Stevens-Linders, P. van Hoogen et al., *Nat. Genet.* **41** (2009) 838.

[20] J. Thusberg, A. Olatubosun and M. Vihinen, *Hum. Mutat.* **32** (2011) 358.

[21] S. Hicks, D.A. Wheeler, S.E. Plon and M. Kimmel, *Hum. Mutat.* **32** (2011) 661.

From Genome Sequence Analysis to Inferring Bacteriophage Infection Strategies

Jelena Guzina [a]
Faculty of Biology, University of Belgrade, Studentski trg 16, Belgrade, Serbia

Marko Djordjević [b]
Faculty of Biology, University of Belgrade, Studentski trg 16, Belgrade, Serbia

Abstract

Bacteriophages are viruses that can specifically destroy a given pathogenic bacteria. Understanding their infection strategies has recently came in focus, particularly in the context of increased resistance of pathogenic bacteria to antibiotics. However, a common way of analyzing bacteriophage infection strategies is labor-intensive and involves a combination of biochemical and bioinformatic approaches and macroarray measurements. We here investigate to what extent one can understand gene expression strategies of lytic phages, by directly analyzing their genomes through

[a] e-mail address: jeca.282@gmail.com
[b] e-mail address: dmarko@bio.bg.ac.rs

bioinformatic methods. We address this question on an example of a recently sequenced lytic bacteriophage 7 – 11 that infects bacterium *Salmonella enterica*. Our main result is identification of novel promoters for the bacteriophage-encoded sigma factor. Identifying the promoters is based on a simple procedure of pairwise alignments of intergenic regions, which correctly identifies the desired motifs; interestingly, standard methods for promoter recognition, which are based on Monte-Carlo procedures, fail to correctly identify the promoters. Furthermore, we also identified promoters for bacterial-encoded sigma factor in the 7 – 11 genome, by using a recently improved model of specificity of bacterial promoters. Identification of all the promoters allows clustering the genes in putative early, middle and late class, which in consequence reveals the bacteriophage infection strategy. We therefore argue that direct analysis of bacteriophage genome sequences is a plausible first-line approach for understanding bacteriophage transcription strategies.

1. Introduction

Bacteriophages represent a group of viruses that is dominant in the microbial world, which to a large degree outnumber the other life forms in the Biosphere [1-2]. While bacteriophage genomes are short, analyzing their genomic sequence is noticeably complicated by genetic exchange, which results in genome mosaicism. As a consequence a large number of genes (almost 80%) in a novel phage does not code for proteins of known functions [1]. On the other hand, bacteriophages share a lot of similar features, like gene expression strategies. For example, genes of a large number of bacteriophages can be designated as the ``early``, ``middle`` and ``late``, based on the temporal pattern of their expression during infection [3].

An additional interest for analyzing bacteriophage gene expression strategies comes with recent occurrence of the bacterial strains resistant to antiobiotics, i.e. due to their relevance for bacteriophage-based therapy treatments. Such successful treatements may use the protein products that bacteriophages express during infection, such as the bacterial RNA polymerase inhibitors, cell wall lysins etc. To understand functions that these molecules perform, the analysis of the bacteriophage gene expression strategy during the infection has arisen as the main objective [4-5]. This approach was up-to now achieved through an approach that heavily relays on experimental measurements, which includes the temporal analysis of gene expression by macroarray measurments, and a biochemical analysis of the promoter elements in the bacteriophage genomic sequence [6].

The previous startegy requires extensive application of time and resources, which is largely impractical given the exponentially growing pace of bacteriophage genome sequencing. That led to the need for developing the more effective methods for acquiring insights into the strategy of bacteriophage infections. An attractive possibility is to extract as much information as possible directly from the bacteriophage genome sequence, by using bioinformatic methods. Exploring this possibility is the main goal of this paper.

Particularly challenging for the analysis are a large number of bacteriophages that express their own RNA polymerase or sigma factor. At the beginning of the life cycle, these viruses use the holoenzyme RNA polymerase – σ factor of a host bacterium for initiating the gene

expression, followed by the repression of this holoenzyme`s activity, which leads to the shut-off of the expression of bacterial genes. The bacteriophage transcription process then swicthes to using its own RNA polymerase/sigma factor in the transcription process, which leads to a completion of the viral gene expression [6].

The key element in understanding such transcription strategies is the prediction of the promoter elements that these, bacteriophage encoded, σ factors/RNA polymerases recognize. These typically show resemblence with RNA polymerases and σ factors of other phages and σA group of sigma factors. However, the level of homology is almost always insuficient for inferring the specificity of promoter elements that they recognize. Consequently, the prediction of promoter elements reduces to a highly non-trivial bioinformatic task, since it comes to a prediction of a few, possibly degenerate, ~10 bp motifs in a 50 – 100 kb long sequence. In line with this, it is often reported that promoters with such organization are hardly detectable by bioinformatic methods divised for such problems, i.e. MLSA (Multiple Local Sequence Alignment) algorithms.

As a model bacteriophage to explore inferring gene expression strategy directly from the genome sequence, we will use recently sequenced bacteriophage 7 – 11. The phage has ~ 90000 bp long double-stranded DNA genome, which infects bacterium *Salmonella enterica*, serovar Newport. No experiments are performed on this bacteriophage, so none of its gene expression control mechanisms are known in advance. Our strategy will therefore be to analyze to what extent one can infer a global view of the bacteriophage gene expression strategy directly from its genome sequence.

On the other hand, the phylogenetic analyses indicate homology of bacteriophage 7-11 with phiEco32 phage that infects bacterium *Escherichia coli*, which was analyzed to detail before [7]. This therefore opens a possibility to use the information on homology with phiEco32 phage, in order to easily validate the obtained bioinformatic predictions.

2. The 7-11 genome arrangement

The total number of 151 ORF was detected in the phage 7 – 11 genomic sequence, 30 of those oriented in the``+``, and the other 121 gene in the ``-`` transcription direction. The genes that possess homologs

in databases mostly infer the homology with the phiEco32 bacteriophage. All the genes are sharply grouped into the ``+`` oriented cluster, which contains the structural and DNA packaging genes, and the ``-`` oriented cluster, composed of the functional genes (Figure 1).

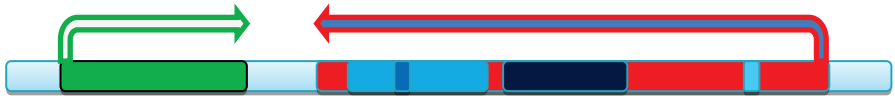


Figure 1. The architecture of the bacteriophage 7 – 11 genome. The two arrows indicate the direction of transcription for the two clasters of genes, so that the red and the green indicate, respectively, the genes with ``-`` and ``+`` transcription orientation. The following colour scheme indicates groups of genes with specific function: dark blue, the genes involved in nucleotide metabolism; blue, the genes involved in DNA and RNA metabolism; blue, the sigma factor gene; magenta, the antisigma factor gene.

The ``-`` cluster genes (the functional genes) can be divided into two different subgroups – genes involved in processes of genome maintainance and expression (DNA replication and transcription) and genes involved in the metabolism of nucleotides. Among the functional genes are also found the genes that code for the ECF family RNA polymerase sigma factor and the antisigma factor. The position of the latter belongs to the most upstream cluster segment. The homology with the phiEco32 coliphage indicates that the possible antisigma factor function is the host RNA polymerase inhibition, which leads to the shut-off of the host and early phage gene transcription.

3. The promoter element detection

Our first-line approach for the phage-specific promoter detection was the usage of MLSA (Multiple Local Sequence Alignment) algorithms. These algorithms represent a standard procedure for detection of promoters that typically appear as degenerated motifs, in a large copy number. The statistical significance determination of the MLSA algorithms` results is still an open problem. The reliability of predictions is validated based on their robustness, that is, the predictions obtained by multiple runs of the same MLSA algorithm should match. The same applies to the usage of different implementations of the basic algorithm (the BioProspector and Gibbs Motif Sampler in our case).

The MLSA algorithms did not yield robust predictions. Therefore, our next assumption was that the phage-specific promoters appear in the 7 – 11 genome as highly conserved motifs in a low copy number, which led to the method based on pairwise alignment of the intergenic regions, through a non-standard usage of BLAST.

By using the BLAST algorithm we detected three putative phage-specific promoters, located upstream from the gene number 1 (Table 1).

The 12 bp long promoter is built of 2 starting bp ``TG``, followed by the core motif ``TGATGT`` and an extra ``TATA`` element. The core motif was further used to track the additional putative promoters, by searching the genomic sequence with the approapriate MATLAB function. Six additional copies were detected upsteram from the genes number 1 (2 copies), 25, 88, 116 and 122. The repeats detected through BLAST are highly statisticallly significant (a low P value). The P value of the 3 large repeats, 12 bp long, in the sequence ~13000 bp long (the length of the intergenic regions) can be estimated at $\sim 10^{-7}$, whereas of the 9 short repeats, 6 bp long, at $\sim 10^{-2}$.

	tAATGT_AtA		
>1	gggggatgtgtgatgttataacataaaggg	2178	
>1	tggtaattatgtgatgttatattgtatcacc	2082	Long repeats with 'TATA'
>1	taaaataacgggatgttataacgtaacaat	1931	
>1	agttgaaggagtgatgtgtaa	2826	
>1	gcttcaccgtttgatgtagtcactatacagc	1176	
>25	agttgaagtgatgtttggggtcgtgaag	10	Short repeats with central
>88	tgtgtttgatgtatctgtaagcatca	8	'TGATGT'
>116	ttcgtgttctttgatgtagactctctctgaa	116	
>122	gattaaagacctgatgtaactatcaagccca	167	

Table 1. The putative promoter elements recognized by phage-encoded sigma factor in the 7 – 11 genome; The sequences are flanked by the number of the downstream gene (left) and the starting coordinate within the intergenic region (right).

Based on the sequence and P value dissimilarities, but also the genome locations, we argue that the two groups of detected elements represent distinct classes of promoters. The long repeats are localized upsteram from structural genes, which means that these elements direct the transcription of ``late`` genes. On the other hand, the short repeats are also found within the downstream segment of the functional cluster, so these elements are possibly engaged in the expression of ``middle`` genes. It is important to emphasize that each one of these motifs is localized downstream with regard to the antisigma factor gene.

Next, we also searched for σA-dependent promoter elements within the bacteriophage genome. The search is based on a simple strategy of detecting sequences with a sufficiently close match to -35 and -10 promoter elements; specifically we allow for up to two mismatches from the consensus elements. This search detected two putative promoters located in the rightmost intergenic region (located in the 5' end of the genome), where both of the promoters have "-" orientation.

4. Comparison of the 7 – 11 phage with the phiEco32

The genomic analysis results of the novel 7 – 11 bacteriophage confirm the assumption on its homology with the phiEco32 coliphage. The detected gene functions, the gene clustering and layout within respective clusters infer a high degree of similarity between the genomic architectures of two phages [7].

Just like the phiEco32, the 7 – 11 phage has genes that code for the sigma and antisigma factors. Based on this, the 7 – 11 phage falls into the group of bacteriophages that code its own RNA polymerase/sigma factor. Related to the topic, it is important to emphasize that the obtained predictions for phage-specific promoters are not only statistically significant, but also similar to the experimentally confirmed phiEco32 promoter.

Furthermore, the 7 – 11 and phiEco32 phages share a similar promoter layout on the genome map. The 7 – 11 genome containes one phage-specific promoter inside of the structural gene cluster with five additional copies right upstream from the cluster, whereas the phiEco32 genome posesses three evenly distributed copies of phage-specific promoters within the respective cluster. Also, both genomes have 3 copies of phage-specific promoters located in the downstream segment of the functional gene cluster, upsteram from the genes involved in the DNA repilcation and nucleotide metabolism, and downstream from the antisigma factor gene. The results of macroarray measurments infer that these 3 promoters direct the transcription of ``middle`` genes in the phiEco32 genome [6]. Based on this fact, our assumption on existence of two distinct classes of phage-specific promoters in 7 – 11 genome can be validated.

Finally, regarding the promoters recognized by bacterial σ^{70}-factor, these elements can be found in both genomes in the large intergenic region on the 5` end in the ``-`` transcriptional orientation, with few additional copies inside of the functional gene cluster [6].

5. Inferring phage 7-11 life cycle

Having in mind the promoter layout on the genome and the clustering of genes, the phage 7 – 11 life cycle can be summarized in the following way: Right upon the bacterial cell infection, the phage intiates the expression of ``early`` genes from the σ^{70}-dependent promoters, by using the host holoenzyme RNA polymerase – σ^{70} factor. By the term ``early`` are considered the upstream functional genes, together with the antisigma factor gene. Because the localization of promoters is upstream from the entire cluster, ``early`` genes are transcribed in a form of a long operon – the feature already detected in a number of other bacteriophages [8].

Upon the expression, the antisigma factor forms a complex with the holoenzyme and abolishes its activity, which results in a termination of the ``early`` phage and bacterial genes transcription. The expression of the remaining genes is, therefore, directed from the promoters recognized by the phage-encoded sigma factor. Interestingly, the gene coding for the phage sigma factor is localized downstream from these promoters, which means that is most probably transcribed from both phage-specific and σ^{70}-dependent promoters. Notice here that the σ^{70}-dependent promoter localization infers the transcription of the entire functional cluster, but also the presence of two copies of the σ^{70}-dependent promoters localized in the upstream vicinity of the phage sigma factor gene. This indicates that the initial amounts of the sigma factor, which are necessary to commence the gene expression from the phage-specific promoters, are transcribed from the promoters recognized by bacterial sigma factor. When these promoters are blocked by the antisigma factor, the sigma factor expression continues from the phage-specific promoters. This is how the sigma factor remains active at the end, like in the beginning of infection. The genes with this transcription pattern are depicted as ``middle`` genes [6].

The last group of genes – the ``late`` ones, consists of structural and DNA packaging genes. Only the promoters localized upstream from these genes have the extra ``TATA`` element, which is likely responsible for their late transcription, having in mind that the ``late`` genes, just like the ``middle`` ones, are transcribed from the phage-specific promoters. After the``late`` genes expression, the bacteriophage 7 – 11 life cycle is

completed, and a large number of virions ready to enter the new infection cycles.

6. Conclusion

In summary, we here analyzed the recently sequenced bacteriophage 7 – 11 genome. The most significant result was the prediction of promoter elements, recognized by both phage-encoded and host bacterial σ factor, directly from the genomic sequence. The identified promoters recognized by the phage-encoded sigma factor show a new promoter specificity, distantly related to the σA family. The new specificity prediction is highly plausible, not only based on the estimated statistical significance, but also on the resemblence with the experimentally confirmed phiEco32 promoter elements. This verifies our approach for direct bioinformatic detection of phage promoters of unknown specificities. The promoter localizations together with standard analysis of the genome sequence (gene and homology predictions) enabled the prediction of the phage 7 – 11 temporal gene expression classes – clustering of the genes into the ``early``, ``middle`` and ``late`` genes. The predicted temporal gene expression classes are consistent with the known functions of the phage 7 – 11 genes. The results presented here strongly suggest that bioinformatic methods may serve as a first-line approach to analyze the life cycles of novel bacteriophages, especially having in mind a large number of sequenced phage genomes and a relatively small number of thoroughly studied representatives [1, 8].

Acknowledgement

This work is partially supported by Marie Curie International Reintegration Grant within the 7th European community Framework Programme (PIRG08-GA-2010-276996) and by the Ministry of Education and Science of the Republic of Serbia under project number ON173052.

References

1. Hatfull, G.F. and R.W. Hendrix, *Bacteriophages and their genomes.* Curr Opin Virol, 2011. **1**(4): p. 298-303.
2. Savalia, D., et al., *Genomic and proteomic analysis of phiEco32, a novel Escherichia coli bacteriophage.* J Mol Biol, 2008. **377**(3): p. 774-89.
3. Hinton, D.M., *Transcriptional control in the prereplicative phase of T4 development.* Virol J, 2010. **7**: p. 289.
4. Haq, I.U., et al., *Bacteriophages and their implications on future biotechnology: a review.* Virol J, 2012. **9**: p. 9.
5. Joerger, R.D., *Alternatives to antibiotics: bacteriocins, antimicrobial peptides and bacteriophages.* Poult Sci, 2003. **82**(4): p. 640-7.
6. Pavlova, O., et al., *Temporal regulation of gene expression of the Escherichia coli bacteriophage phiEco32.* J Mol Biol, 2012. **416**(3): p. 389-99.
7. Kropinski, A.M., E.J. Lingohr, and H.W. Ackermann, *The genome sequence of enterobacterial phage 7-11, which possesses an unusually elongated head.* Arch Virol, 2011. **156**(1): p. 149-51.
8. Djordjevic, M., et al., *Quantitative analysis of a virulent bacteriophage transcription strategy.* Virology, 2006. **354**(2): p. 240-51.

Kink Solitons and Breathers in Microtubules

Slavica Kuzmanović [a]
University of Priština, Kosovska Mitrovica, Serbia

Abstract

We present two new dynamical models of microtubules (MTs). Microtubules are major part of cytoskeleton and serve as a network for motor proteins. They are holly cylinders formed by 13 long structures called protofilaments (PFs). Elementary units of PFs are 8nm long electric dipoles called dimers. Known models of MTs assume only one degree of freedom per dimer. According to the chosen coordinate the models can be called as longitudinal and radial. We start from the MT Hamiltonians from which nonlinear partial differential equations describing MT evolution are derived. Basically, two analytical approaches for solving these nonlinear equations are known. These are developed on the continuum and semi-discrete approximations. According to the first one kink and antikink solitons move along PFs. The second approach brings about localized modulated waves usually called breathers.

1 Introduction

Microtubules (MTs) are important cell proteins. They represent cytoskeleton and serve as a road network for motor proteins (kinesin and dynein)

[a] e-mail address: slavica.kuzmanovic@pr.ac.rs

dragging different molecular cargos. Their lengths may span dimensions from the order of micrometer to the order of millimeter. Constituent parts are electric dipoles called dimers. MTs are unstable systems with very complicated structure. To be able to study these systems biophysicists u assume only one degree of freedom per dimer, either longitudinal assume only one degree of freedom per dimer, either longitudinal [1,2] or radial [3]. This allows researchers to obtain partial differential equations describing MTs complex nonlinear dynamics. The solutions can be under certain conditions obtained by analytical procedures, while the numerical methods allow the full treatment of the mentioned problem. The first nonlinear model of MTs was introduced twenty years ago [4].

There are usually 13 longitudinal PFs covering the cylindrical walls of MTs, as shown in Fig. 1. Each PF represents series of proteins known as tubulin dimers. Each dimer is an electric dipole whose whose length is $l = 8$nm and width is $d \approx 4$nm. The constituent parts of the dimers are α and β tubulins, corresponding to + and - side, respectively. It might be interesting to point out that MTs are always oriented with positively charged ends towards a cell nucleus, and with a negatively one towards a membrane. Notice that positively charged end corresponds to biologically minus end and vice versa. Namely, biologically positive end is attributed to more intensive growing end during its polymerization.

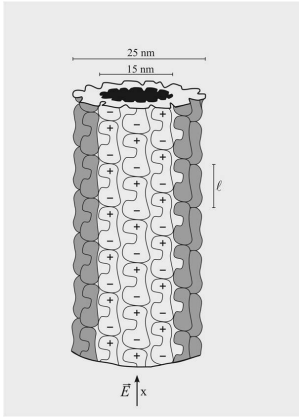


Figure 1: A structure of a microtubule.

2 Radial (φ-model)

The expression for the Hamiltonian of one PF is

$$H = \sum_n \left[\frac{I}{2} \dot{\varphi}_n{}^2 + \frac{k}{2} (\varphi_{n+1} - \varphi_n)^2 - pE \cos \varphi_n \right], \tag{1}$$

which, after a expansion of the cosine function, becomes

$$H = \sum_n \left[\frac{I}{2} \dot{\varphi}_n{}^2 + \frac{k}{2} (\varphi_{n+1} - \varphi_n)^2 - pE + pE \left(\frac{\varphi_n{}^2}{2} - \frac{\varphi_n{}^4}{24} \right) \right]. \tag{2}$$

Here dot means a first derivative with respect to time, I is a moment of inertia of the dimer and the integer n determines a position of the dimer in the PF. In equation (2), p is the electric moment and E is an electric field. It is obvious that the first term in equation (2) represents a kinetic energy, while the second one is a potential energy of the chemical interaction between the dimers belonging to the same PF. The interaction between the neighboring dimers belonging to different PFs is much smaller and the corresponding energy is neglected.
Dynamical equations of motion, if a viscosity momentum $M_v = -\Gamma \dot{\Phi}_n$ (Γ is a viscosity coefficient) is incorporated into this equation, we obtain the following form

$$I\ddot{\varphi} - k(\varphi_{n+1} + \varphi_{n-1} - 2\varphi_n) + pE \sin \varphi_n + \Gamma \dot{\varphi_n} = 0. \tag{3}$$

Therefore, using (3) and a transformation

$$\varphi = \psi\sqrt{6}, \tag{4}$$

we obtain

$$\frac{I}{pE}\ddot{\psi}_n - \frac{k}{pE}(\psi_{n+1} + \psi_{n-1} - 2\psi_n) + \psi_n - \psi_n{}^3 + \frac{\Gamma}{pE}\dot{\psi}_n = 0. \tag{5}$$

Using the continuum approximation for the equation (5), and travelling wave ansatz

$$\xi = (x - vt) \Rightarrow \psi = \psi(x,t) = \psi(\xi), \tag{6}$$

we can easily obtain the following ordinary differential equation

$$\alpha \frac{d^2\psi}{d\xi^2} - \rho \frac{d\psi}{d\xi} + \psi - \psi^3 = 0, \tag{7}$$

where α and ρ are parameters: $\alpha = \frac{I\omega^2 - kl^2}{pE}$, $\rho = \frac{\omega\Gamma}{pE}$, while $\frac{d^2\psi}{d\xi^2}$ and $\frac{d\psi}{d\xi}$ are derivatives with respect to ξ. A solution of this ODE will describe nonlinear dynamics of MTs.

3 Longitudinal (u-model)

The Hamiltonian for one PF is represented as follows for u - model

$$H = \sum_n \left[\frac{m}{2}\ddot{u}_n^2 + \frac{k}{2}(u_{n+1} - u_n)^2 + V(u_n) \right], \tag{8}$$

where dot means the first derivative with respect to time, m is mass of the dimer, k is an intra-dimer stiffness parameter and the integer n determines the position of the considered dimer in the PF. The first term represents a kinetic energy of the dimer, the second one is a potential energy of the chemical interaction between the neighboring dimers belonging to the same PF and the last term is the combined potential

$$V(u_n) = -qEu_n - \frac{1}{2}A{u_n}^2 + \frac{1}{4}B{u_n}^4. \tag{9}$$

It is obvious that the nearest neighbor approximation is used. However, this does not mean that the influence of other neighboring PFs is completely ignored as the value of the electric field E depends also on the dipoles belonging to them. We can straightforwardly obtain an appropriate dynamical equation of motion for u - model. The viscosity of the solvent is taken into consideration too. This is done by introducing viscosity force $F_v = -\gamma\dot{u}$ into the dynamical equation of motion (γ is viscosity coefficient). All this brings about the following nonlinear partial differential equation

$$m\ddot{u}_n - k(u_{n+1} + u_{n-1} - 2u_n) - qE - Au_n + B{u_n}^3 + \gamma\dot{u}_n = 0. \tag{10}$$

By introducing dimensionless function ψ

$$u = \sqrt{\frac{A}{B}}\psi, \tag{11}$$

and after substituting equation (11) in (10) we have

$$\frac{m}{A}\ddot{\psi}_n - \frac{k}{A}(\psi_{n+1} + \psi_{n-1} - 2\psi_n) - \psi_n + {\psi_n}^3 + \frac{\gamma}{A}\dot{\psi}_n - \sigma = 0. \tag{12}$$

The parameter σ is proportional to the internal electric field strength $\sigma = \frac{qE}{A\sqrt{A/B}}$. For u - model we obtain ODE from the equation (12)

$$\alpha\frac{d^2\psi}{d\xi^2} - \rho\frac{d\psi}{d\xi} - \psi + \psi^3 - \sigma = 0, \tag{13}$$

which contains the following new parameters $\alpha = \frac{m\omega^2 - kl^2}{A}, \rho = \frac{\gamma\omega}{A}$.

4 Kinks solution for (φ-model) and longitudinal (u-model)

Equation (7) can be solved using the modified extended tanh-function (METF) method. According to this procedure we look for the possible solution of the form

$$\psi = a_0 + \sum_{i=1}(a_i\Phi^i + b_i\Phi^{-i}), \tag{14}$$

where the function $\Phi = (C\xi)$, is a solution of the well known Riccati equation

$$\frac{d\Phi}{d\xi} = b + \Phi^2. \tag{15}$$

The parameter b is a real constant and $\frac{d\Phi}{d\xi}$ is the first derivative. Using modified extended tanh-function (METF) method, we obtain expected solution for φ

$$\varphi(x,t) = \frac{\sqrt{6}}{2}\left[1 + \tanh\left(\frac{3}{4\rho}\xi\right)\right]. \tag{16}$$

When kink soliton reaches the positive end of MT the dimers rotate for large angle in the direction from the main axes, for $\sqrt{6}$ towards outside, causing well known blowups of MT. Depolymerization always starts from the biologically positive end, i.e. negatively charged end. Biological equivalence of the instability is the blowups of MT from stable $\varphi(-\infty) = 0$ to unstable state $\varphi(+\infty) = \sqrt{6}$.
Using modified extended tanh-function (METF) method, we obtain expected solution for u - model

$$\psi_i(\xi) = a_{0i} - \sqrt{1 - 3{a_{0i}}^2}\tanh\left(\frac{3a_{0i}}{\rho}\sqrt{1 - 3{a_{0i}}^2}\xi\right). \tag{17}$$

There are three solutions, because $i = 1, 2, 3$. ψ_1 is one of them.
According to previous equations we find one kink and antikink solitons moving along PFs. An example, coming from the longitudinal model, is shown in Fig. 2. The function $\psi_1(\xi)$ is shown for two values of the parameter σ and ρ. Obviously, this is an anti-kink soliton. The parameter ρ affects its slope and width of the soliton, but not the character of solution. It is obvious that the solitonic width is proportional to viscosity. In equation (17), $a_{01} = \frac{1}{2\sqrt{3}}$ is the parameter. The coordinate $\xi = kx - \omega t$ is introduced to switch from the partial to the ordinary differential equation.

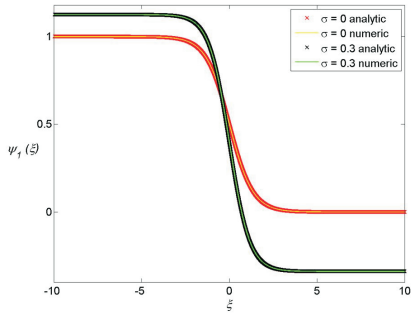


Figure 2: The function $\psi_1(\xi)$.

5 Semi-discrete approximation

The semi-discrete approximation [5-7] starts by introducing small fluctuations

$$\psi_n = \epsilon\phi_n; (\epsilon << 1), \tag{18}$$

substituting equation(18) in (5), but neglect the last member of the equation, we obtain

$$\frac{I}{pE}\ddot{\Phi}_n - \frac{k}{pE}(\Phi_{n+1} + \Phi_{n-1} - 2\Phi_n) + \Phi_n - \epsilon^2\Phi_n^3 + 0(\epsilon^3) = 0. \tag{19}$$

We expect the following solution

$$\Phi_n(t) = F(\xi)e^{i\theta_n} + \epsilon F_0(\xi) + cc + 0(\epsilon^2), \tag{20}$$

where the functions $F(\xi)$ and $F_0(\xi)$ are to be determined. Using the the continuous mode approximation $nl \to z$ and introducing new coordinates

$$Z = \epsilon z; T = \epsilon t. \tag{21}$$

Should be determined after straightforward procedure we obtain dispersion relation $\omega^2 = \frac{4k\sin^2(ql/2)+pE}{I}$, and group velocity of travelling waves $V_g = \frac{lk}{I\omega}\sin(ql)$.

By introducing new variables, and then trimming some members using the dispersion equation P, we obtain equation with the dispersion coefficient $P = \frac{1}{2\omega}\left[\frac{l^2k}{I}\cos(ql) - V_g^2\right]$, and the coefficient of nonlinearity $Q = \frac{3pE}{2I\omega}$. Finally the system equation is in a form of the nonlinear Schrödinger equation

$$iF_\tau + PF_{ss} + Q\left|F\right|^2 F = 0. \tag{22}$$

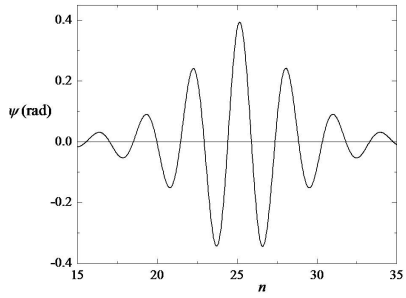


Figure 3: The function ψ_n for $t = 3$ns, presenting the localized modulated soliton (breather)

Solution of nonlinear Schrödinger equation can be written in a form

$$F(S,\tau) = A_0 \left(\frac{S - u_e\tau}{L_e} \right) exp \frac{iu_e(S - u_c\tau)}{2P}, \tag{23}$$

where $u_e > 2u_c$, $S = Z - V_gT$ and $\tau = \epsilon t$.

The semi-discrete approximation brings about localized modulated waves usually called breathers. An example, describing the radial model, is shown in Fig. 3.

$$\psi_n(t) = 2A\, sech \left(\frac{nl - V_e t}{L} \right) cos\left(\Theta nl - \Omega t\right), \tag{24}$$

where amplitude $A \equiv \epsilon A_0 = U_e\sqrt{\frac{1-2\eta}{2PQ}}$, width of soliton $L \equiv \frac{L_e}{\epsilon} = \frac{2P}{U_e\sqrt{1-2\eta}}$, $\Theta = q + \frac{U_e}{2P}$, $\Omega = \omega + \frac{(V_g+\eta U_e)U_e}{2P}$ and $V_e = V_g + U_e$.
Coherent metod $V_e = \frac{\Omega}{\Theta} \Rightarrow U_e = \frac{P}{1-\eta}\left[-q + q\sqrt{1 + \frac{2(1-\eta)}{Pq^2}(\omega - qV_g)}\right]$,
where $0 \leq \eta < 0.5$, $\eta = \frac{u_c}{u_e}$ and $U_e = \epsilon u_e$.

6 Conclusions

In this article two new models describing nonlinear dynamics of MTs are introduced. We show that these models bring about the equations whose solutions have both stable and unstable asymptotes corresponding to the two edges of the PF. MTs serve as a road network for motor proteins

(kinesin and dynein) dragging different cargos such as vesicles and mitochondria. A soliton generated in MTs is a signal which activates a proper motor. According to developed models MT is dynamically very unstable structure which corresponds to the reality. The MT life time in normal cells is about 2-4 hours while its depolymerization (disintegration) occurs in a few seconds.

Acknowledgements

Author gratefully acknowledges for inspiring discussions with Dr. Slobodan Zdravković.

References

[1] S. Zdravković, L. Kavitha, M. V. Satarić, S. Zeković, J. Petrović, *Chaos, Solitons Fract.* **45** (2012) 1378.

[2] S. Zdravković, M. V. Satarić, S. Zeković, *Europhys. Lett.* **B 102** (2013) 3802.

[3] S. Zdravković, M. V. Satarić, A. Maluckov, A. Balaž, *Appl.Math.Comput.* (2014), http://dx.doi.org/10.1016/j.amc.2014.03.113.

[4] M. V. Satarić, J. A. Tuszyński and R. B. Žakula, *Phys. Rev.* **48** 589 (1993).

[5] M. Remoissenet, *Phys. Rev.* **B 33** (1986) 2386.

[6] T. Kawahara, *J. Phys. Soc.* Japan **35** (1973) 1537.

[7] S. Zdravković, A. N. Bugay, Localized modulated waves in microtubules. Work in preparation.

Protein Subunit Association: NOT a Social Network

Mounia Achoch [a]
LISTIC,University of Savoie, Annecy le Vieux, France
Giovanni Feverati [b]
LAPTH UMR 5108, University of Savoie, CNRS, Annecy le Vieux, France

Laurent Vuillon [c]
LAMA UMR 5127, University of Savoie, CNRS, Le Bourget du Lac, France

Kave Salamatian [d]
LISTIC, University of Savoie, Annecy le Vieux, France
Claire Lesieur [e]
AGIM FRE 3405, University of Grenoble Alpes, CNRS, Grenoble, France

Abstract

Most proteins cannot function as single unit but associate subunits via the formation of protein interfaces, to be biologically active. How the amino acids involved in subunit association, so-called hot spots, regulate the formation of a protein interface is still an open question. Here, we show how network and graph

[a] e-mail address: Mounia.Achoch@univ-savoie.fr
[b] e-mail address: feverati@free.fr
[c] e-mail address: laurent.vuillon@univ-savoie.fr
[d] e-mail address: kave.salamatian@univ-savoie.fr
[e] e-mail address: claire.lesieur@agim.eu

theories can help addressing the role of hot spots. We built a MatLab code called SpectralPro which identifies hot spots and reconstructs the protein interface as a subnetwork of hot spots in interaction, with the hot spots as nodes and the bonds between hot spots as links. Using as a case study, the cholera toxin B pentamer (five subunits), we investigate if the degree of a node, namely the number of contacts of a hot spot, is important in the formation of an interface. The degree of a node is known to be important in many real networks. For example in social networks, hubs control the communication between most nodes and as such are vulnerable to changes. But our result shows that in the toxin interface sub-graph hub-like nodes are less vulnerable to change than single link node.

1 Introduction

Proteins are biological entities made of a chain of amino acids bound to one another in a specific order, called the primary structure or the amino acid sequence of the protein. Based on the sequence and the environment, the protein acquires a tridimensional shape called tertiary structure (3D-structure), suitable for its biological function. The set of reactions leading to the functional 3D-structure is the folding of the protein. It involves the formation of bonds/interactions between atoms of the amino acids of a single chain. These interactions are called intramolecular amino acid interactions. There exist proteins which function as oligomers by associating several copies of the same chains (homo oligomers) or of different chains (hetero oligomers). The association of chains forms the quaternary structure (4D-structure) of the proteins. The zone of contact between two associated chains is called the protein interface. The protein interface involves the formation of interactions/bonds between atoms of the amino acids of adjacent chains. These interactions are called intermolecular amino acid interactions. Among the amino acids involved in intermolecular amino acid interactions, only a subset is important for the formation of the interface, those are called hot spots [1].

Some protein oligomers are involved in diseases as virulence factors, like the notorious cholera toxin responsible for the cholera disease [2]. Understanding and predicting how such proteins assemble into oligomers is essential for designing appropriate inhibitors capable of preventing their pathological assemblies. The design of such inhibitor entails to identify the hot spots and understand their role in the formation of an interface.

There are numerous algorithms capable of identifying hot spots from the 3D structure of protein oligomers whose atomic coordinates are available from the Protein Data Base (www.rcsb.org/pdb/). However, these algorithms do not provide means to understand how the hot spots orchestrate the formation of an interface. We propose to consider hot spots as nodes and bonds between hot spots as links, and to build a subgraph or a subnetwork of hot spots in interaction to model the interface. Sub graph because it describes only a local feature of the protein chain, namely the interface and not the entire chain, which would be a graph. The hot spots can be distinguished by network measures and we can look for correlation between the network's measures and the importance of the hot spots in terms of interface formation. A good overview of network measures can be found in [3]. Our case of study is the cholera toxin B subunit pentamer ($CtxB_5$) produced by *Vibrio cholera*. We have written a Matlab code that reasonably identifies the hot spots of the $CtxB_5$'s interface and builds a sub-graph of the toxin's interface based on a matrix of contacts. We look if the degree of the nodes, namely the number of contacts of the hotspots, has any relevance in terms of the formation of the toxin's interface.

2 Methods

SpectralPro. SpectralPro uses the Cartesian coordinates of the atoms of the 3D-structure of $CtxB_5$ as an input. These coordinates can be extracted from the PDB under the PDB code 1EEI. Each chain of the pentamer is considered as a set of points in the space whose positions are the Cartesian coordinates (x, y, z) of the atoms of the chain. The atoms of the chain 1 constitute the set 1 (S1), the atoms of the chain 2, the set 2 (S2) and the atoms of the chain 5, the set S5. SpectralPro calculates distances between every atom of S1 and every atom of the four other sets (interchain distances) but ignores the distances between atoms of a single set (intrachain distances). It chooses for every atom the 10 closest atoms and among these, it selects the pairs of atoms distant of a maximum of 5 Angstrom. Every atom is involved in a certain number of pairs, namely it has a certain numbers of contacts. SpectralPro builds a N x N matrix with the selected intermolecular atoms as the nodes N and the elements of the matrix as their number of contacts. SpectralPro also builds a coarse-grained matrix where the atoms are replaced by their respective amino acids as nodes. A weightless matrix is produced where the elements of the matrix are one when the amino acids have at least one pair of atoms in contact and zero when they don't. The weightless matrix provides for every amino acid, its

number of amino acid contacts.

Fold X. The effect of a local change (amino acid mutation) on the formation of the toxin interface is measured by generating a virtual single point mutation on the toxin PDB with Fold X and by calculating the free energies of interactions at the interface for the non mutated (wild-type) and the mutated proteins [4]. The difference between the two energies measures the effect of the mutation. The amino acid plays a role in the formation of the interface if its mutation leads to a non zero energy difference.

3 Results and discussion

The goal of the investigation is to develop an appropriate tool to reconstruct the $CtxB_5$ interface as a sub-graph of hot spots in interaction, analyze some graph properties to determine their relevancy in terms of the toxin assembly.

3.1 Identification of hotspots

The first step is to test if SpectralPro is capable of identifying hot spots. The details on how SpectralPro detects amino acid in contact is described in the methods. Because SpectralPro reads the atoms following the amino acid sequence of the chain and selects the closest atoms, it retraces a good reading of the geometry of the two surfaces that make the interface compared to a selection based simply on a cut-off distance. The cut-of distance at 5 Ansgtrom applied subsequently allows to choose the bonds the most chemically probable. It is unlikely that every atom makes ten chemical bonds (ten closest atoms), but the ten links provide a density of interactions instead of evaluating an exact number of interactions. The idea is to obtain an estimate of a probability of interactions of the amino acids. The coarse-grained amino acid sub-graph is built on a square matrix having as rows and columns the amino acids, ordered according to their location along the sequence. The elements of the matrix at position i, j have a one entry if the i-th and j-th amino acids have at least one pair of atoms in interaction (weightless sub-graph).

The sub-graph of the atoms in interaction over the five interfaces of the pentamer has 1498 nodes and 2830 links. In other words, the sub graph is made of 1498 atoms with 2830 closest atoms. The coarse-grained sub-graph of the amino acids in interaction has 283 nodes and also 2830 links (weighted sub-graph). Thus on average every atom has two closest atoms located within 5 Angstrom distance and every amino acid has about five atoms involved in a pairwise interaction. If a single link is counted for every

pair of amino acids, the (weightless) sub-graph has 283 nodes and 422 links. To have an idea of the order of magnitude of a protein interface sub-graph, it is interesting to compare with the world wide web which has 200 million nodes (webpages) and 1.5 billion links, links between two pages.

The amino acids selected as in interaction by SpectralPro are compared to the detection of hot spots by three other available programs (not shown). SpectralPro identifies 283 amino acid contacts over 5 interfaces, with an average of 57 ± 1 hot spots per chain. If we consider the set S5, namely the chain E, SpectralPro identifies 56 hot spots against 39, 57 and 54 for Gemini, PSIBASE and SCOWLP, respectively. Gemini detects hot spots by selecting the mutually closest atoms yielding a more stringent selection than SpectralPro and less hot spots identified [5]. All hot spots detected by Gemini are identified by SpectralPro. PSIBASE as SpectralPro calculates the Euclidean distance to determine pairs of interactions [6]. SpectralPro identifies all the hot spots identified by PSIBASE expect three, making about 5 % false negative. Only one amino acid detected by SpectralPro is not detected by PSIBASE, making less than 2 % false positive. On average in PSIBASE, every hot spot has 5 atoms involved in a pairwise interaction as observed for SpectralPro. SpectralPro identifies all the hot spots identified by SCOWLP expect one, making less than 2 % false negative. There are three amino acids detected by SpectralPro but not by SCOWLP, making about 5 % false positive. SCOWLP identifies pairwise interactions using Eucledian distances and shape-based algorithms [7]. Globally the amino acids selected as hot spots by SpectralPro are consistent with those identified by other programs, supporting that SpetralPro detects hot spots reasonably.

3.2 The degree measure

On a previous study on a large dataset of 1048 interfaces involving the interactions between two beta -strands, we had measured the degree of the nodes of the sub-graph interfaces and looked at the degree distributions [8]. The sub-graphs were built with a different algorithm, called Gemini which selects only a framework of interactions, as mentioned above. The result indicates an exponential degree distribution, no hubs and many nodes with one to three contacts. We have determined statistically that the only amino acids with more than three contacts are R, Y, L and W.

Now we look whether this result is confirmed using SpectralPro which sets less stringency on the selection of hot spots and the number of contacts. The average number of contacts $\overline{k}$ over the five $CtxB_5$ interfaces is 3.1 ± 1.8.

Thus even with SpectralPro, the average number of contacts per residues remains around three.

The degree distribution *P(k)* is the number of hot spots with k degree plotted against the degree k. *P(k)* is calculated for each of the five interfaces of $CtxB_5$ and the average degree distribution and standard deviation is plotted against the degree (Figure 1).

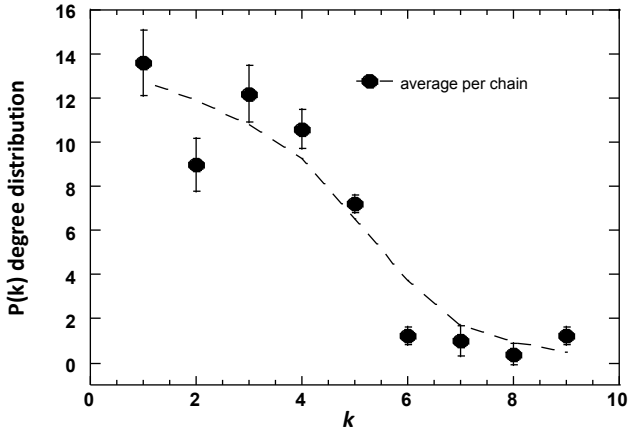


Figure 1: Degree distribution

P(k) for the sub-graph of the $CtxB_5$ interface follows a bell like shape which corresponds to a random network with no hubs but nodes with few links. Again this confirms the observation made on the dataset using Gemini that interface subnetworks do not follow power law degree distribution and have no hubs.

At most the hot spots have 9 contacts, and there are only two such nodes, the amino acid arginine 67 (Arg67) and the leucine 31 (Leu31). Thus the bigger ratio between the highest and lowest degree in the subgraph is 9. On a subgraph of the WWW of 325 729 nodes, which follows a power law degree distribution, the average $\overline{k}$ is 5.46, the ratio between the lowest and highest node degree is 10000. So the hot spots with 9 contacts might be better referred to as hub-like rather than hub. Interestingly, in

comparison the average degrees $\overline{k}$ of the two networks appear rather similar, illustrating the difficulty in interpreting average $\overline{k}$ values for different types of degree distribution. This is discussed in [9].

3.3 Influential nodes

We then explore if the degree of the nodes is any relevant to the formation of the toxin interface. For this purpose, a hot spot with a single contact, lysine 69 (Lys69) and a hub-like hot spot, Arg67 are virtually mutated to an asparagine (Asp, N) using Fold X [4]. The free energy of interaction at the interface is calculated for the mutant and the wild type (WT) proteins. The effect of the mutation is measured as the difference between the wild type and mutant free energies of interaction at the interface. Differences not equals to zero indicate that the mutated hot spot is involved in the formation of the interface. Asparagine is chosen because it has "average" amino acid properties, so if a mutation has no effect on the free energy of interaction, it indicates that the mutated hot spot has average property for the formation of the interface and is plastic to mutation. If a mutation has an effect, the mutated hot spot must have an involvement in the formation of the interface above average, this hot spot can be considered more influential for the formation of the interface and less plastic to mutation. The WT, Lys69Asp and Arg67Asp free energies of interaction are -13,35; -19, 65 and -16, 65 kcal/mol, respectively, as determined by Fold X. This shows that the hot spots are not equally important for the formation of the interface, suggesting their different roles. The free energy of the interface has decreased by a factor of 0.4 and 0.2 upon mutation of the Lys69 and Arg67, respectively. The largest mutational effect on the free energy is for the Lys69Asp mutant over mutation of all other amino acids of the toxin (not shown). Thus the mutation of the single link hot spot Lys69 has more effect on the interface than the mutation of the hub-like Arg67. Thus in contrast to social networks and other real networks, in the sub-graph of the toxin interface, the influence of a node is not directly linked to its degree. More precisely, hub-like residues are not more vulnerable to change, namely mutation, than single link node.

4 Conclusion

In conclusion, we can say that protein interface subnetworks have very different scales compared to other real networks, much less links, lower ratio high degree/low degrees, no hub and behave rather like a random network.

Thus to infer "biological rules", such as the mechanism of assembly or the formation of interfaces, one cannot simply use the network measures that regulate other real networks (www or social network). Intuitively, we could have expected that hub-like hot spots would have been the most influential for the formation of the interface and highly susceptible to mutation as demonstrated for other real networks [10], but that is not the case . Here the result shows that connected does not imply influential in the case of protein interface networks. It remains to be established what makes a node influential if not its degree and to analyze the effect of the mutation on the network.

Acknowledgements

This work is supported by the Federation de recherche of France, FR2914, MSFI (Modelization, Simulations, Fundamental Interactions).

References

[1] Clackson T, Wells JA, *Science* 267(5196):383-6 (1995)

[2] Hirst TR, *J. Moss BI, M. vaughan and A. t. Tu, editor. New York: M. Dekker* 123-84 (1995)

[3] Barabasi A.L, Oltvai Z.N, *Nature reviews Genetics* Feb;5(2):101-13 (2004)

[4] Guerois R, Nielsen JE, Serrano L, *Journal of molecular biology* 320(2):369-87 (2002)

[5] Feverati G, Lesieur C, *PloS one* 5(3):e9897 (2010)

[6] Gong S, Yoon G, Jang I, Bolser D, Dafas P, Schroeder M, et al, *Bioinformatics* May 15;21(10):2541-3 (2005)

[7] Teyra J, Doms A, Schroeder M, Pisabarro MT, *BMC Bioinformatics* 7:104 (2006)

[8] Feverati G, Achoch M, Vuillon L, Lesieur C, *PloS one* in press (2014)

[9] Newman ME, Strogatz SH, Watts DJ, *Phys Rev E Stat Nonlin Soft Matter Phys* Aug;64(2 Pt 2):026118 (2001)

[10] Albert R, Jeong H, Barabasi AL, *Nature* Jul 27;406(6794):378-82 (2000)

From Genetic Code toward Spacetime Geometry

Nataša Ž. Mišić [a]
Lola Institute, Kneza Višeslava 70a, Belgrade, Serbia

Abstract

Numerous arithmetical regularities for the nucleon numbers of the canonical amino acids for quite different systematizations of the genetic code, dominantly based on the *decimal number* 037, indicate the hidden existence of a more universal background regulating system. Generalization of the number 037 reveals the infinite number set whose elements are characterized by *self-similarity*. Here we show that these numbers are uniquely linked to the numeration process, what for the number 037 reveals its unique relation with the golden means, Φ and ϕ. Modification of Φ, ϕ-polynomials gives the *irrationals*, Ψ *and* ψ *, which integrate the self-similarity properties and the scaling by powers of 10*. Moreover, we show that 037 is the simplest case of integer approximation of Ψ, as well as that they enable a simple and an accurate enough derivation of the fine structure constant. These results are consistent with the *Dirac-Eddington large-number coincidence and a quantum gravity theory of the causal dynamical triangulations*. Information processing in Nature realized on such principles would lead to the nested codes and the geometrically based "computation" with a further consequence of fractal structural and dynamical organization, which omnipresence is recognized both in physical and biological realm.

[a] e-mail address: nmisic@afrodita.rcub.bg.ac.rs

1. Introduction

A question of the link between *information* and *emergent phenomena*, considered also as a revival of antique cosmologies based on a dynamic dualism like Plato's *World of Being* (*Forms*, *Ideas*) and *World of Becoming* and Aristotle's *Potential* (*cp. Aperion*) and *Act* (*Aphorismenon*), is a very actual and prospective problem both in physics and biology.

In physics, the consideration of information as an ultimate source of physical reality mainly arises from an effort to unify the microcosm of quantum field theory (Fermi scale [1]) and the macrocosm of general relativity (Hubble scale) into a unique quantum theory of gravity. Despite many approaches to the quantum gravity [2], their most common property is a quantization of gravity at a deeper and a single super unification scale (Planck scale) where is settled a realm of quantum bits. Most approaches interpret relationship between two above physical theories as the *holographic duality* [3], also referred as the *AdS-CFT* (or *Maldacena*) *correspondence* and in a limited sense as *gauge-gravity duality* [4], which ultimately suggests the emergence of gravity and spacetime geometry [5,6], as well as particles and fields [1,7] from informational-entropy theory. Wheeler [8] summarized this idea of the informational origin of particles, fields and spacetime continuum in the famous phrase *It from bit*.

In biology, the emergent phenomena related to the evolution of novelty, integrity and complexity are so obvious and inherent that the *emergence*[1] is actually originated[2] nearly a century and half ago from the life phenomena [9]. On the other side, information becomes a fundamental ingredient of the biological world during development of biophotonics [11], quantum biology [12], consciousness studies [13,14], bioinformatics and related areas. These studies place the *biological coding* at the very heart of the biological organization, as well as its origin and evolution, and ultimately lead to the comprehension that the *organic (biological) codes can be considered as the basic mechanism of macroevolution* [15] and thus to the biological emergence from information. The fact that the biological coding is the most efficiently based on a *nested principle* [16] emphasizes the importance of the first biological code – *genetic code* [17], not only as the origin of life, but

[1] The two types of emergence is differing: *weak emergence* which describes the novel properties of a system (the system's higher level properties) as *reducible* to the underlying properties of system's constituent parts (the system's lower level properties) [10], and *strong emergence* which is in the same sense *irreducible*.

[2] The roots of ideas about emergence found already in Aristotle: "Whenever anything which has several parts is such that the whole is something over and above its parts, and not just the sum of them all, like a heap, then it always has some cause" [*Metaphysics* 1045a 10f].

also as the link between the physical and biological realm with information as an ultimate unifying concept.

Our approach to this objective is based on Shcherbak's [18,19] astonishing discovery of various arithmetic regularities inside the genetic code, wherein we do not look for a specific hypothetical organelles that working as biocomputers inside an organism, but for a universal underlying hypothetical system that causes and controls such arithmetic regularities and which is established beyond an organism.

2. The genetic code and the self-similar numbers

The initial Shcherbak's [18] key result is revealing arithmetic inside the universal genetic code based on *decimal number* 37 *as a nucleon packing quantum of genetic code constituents* – the canonical amino acids and nucleotide bases, with remark that the number 37 is unique in decimal system in the sense that its *three* digit multiples remain multiples modulo 9 by *cyclic permutations* (Fig. 2D) and that similar numbers also exist in some other number systems (13_4, 27_7, 49_{13}).

Shcherbak [18,19] pointed out a variety of different nucleon arithmetic regularities, including those for the free form amino acids and peptide bonded amino acids (the standard block residues and the ionized and protonated side chains), for the compressed, life-size, and split representation of genetic code, for Rumer's and Gamow's division of genetic code, and many other regularities. As an explanation of the arithmetical regularities based on number 37, Shcherbak [19] suggested that a very simple divisibility rule of 37 for base 10 "…simplifies molecular machinery and facilitates the computational procedure of hypothetical organelles working as biocomputers", as well that this divisibility "…criterion is valid for the $PQ = \langle 1_n 1_{n-1} \cdots 1_1 \rangle / n$, if the condition $(q - 1)/n = \mathrm{Int}$ is applied". But our hypothesis was that these numerous arithmetical regularities of nucleon numbers for the genetic code constituents for the quite different systematizations [18-23], dominantly based on decimal number 37, indicate the *hidden existence of a more universal (pre)ordering principle settled in the spacetime geometry* [23]. Motivation for this hypothesis, we found in the unique properties of number 37 (since the essential properties of number 37 are manifested in three-digit notation, hereinafter it will be denoted by 037 [21]).

We showed [22] that the basic property of 037 is a *cyclic equivariability,* vis. *equidistant cycling digit property* (both for the *multipliers* and *digits equidistance*), and that its generalized numbers, the *Shcherbak numbers*[3] ($\mathcal{S}$), have a simple form

$$\mathcal{S}_n(q) = \frac{R_n(q)}{n}, \text{ for } n|q-1, \tag{1}$$

where $R_n(q) = \sum_{i=0}^{n-1} q^i = 11\cdots 11_q$ are the *generalized Niven (Harshad) repunits* of length n in the numeral system of base q such $n, q \in \mathbb{N}_{\geq 2}$ (we assume $\mathbb{N}_{\geq 0} = \mathbb{N}$).

Since for a positive integer m when $m|n$ then $R_m|R_n$, follows that the *irreducible* $\mathcal{S}$ have a form $\mathcal{S}_p(q) = \frac{R_p(q)}{p}$ where p is prime. The condition $p|R_p(q)$ is equivalent to $q^p \equiv 1 \pmod p$ what with Fermat's Theorem $q^{p-1} \equiv 1 \pmod p$ for $(p, q) = 1$ gives $p|q-1$ as the necessary and sufficient condition for the existence of irreducible $\mathcal{S}$ and can be extended to $n|q-1$ for each $\mathcal{S}$, Eq. (1).

The properties of $\mathcal{S}$ follow from the peculiarities of R_n related to its *invariability under the cyclic transformation of discrete sets*. In an arithmetical sense, R_n is an elementary case of cyclic equivariability of the n-digit number, while in a geometrical sense, R_n is an elementary case of discrete rotational symmetry of the n^{th} order. Namely, a repunit satisfies

$$R_n(q) = \frac{q^n-1}{q-1} = \prod_{d|n, d>1} \Phi_d(q), \tag{2}$$

where $\Phi_d(q)$ is the d^{th} *cyclotomic polynomial* which roots lie on the unit circle in the complex plane. Thus R_n relation to the d^{th} roots of unity (for each $d|n$) implies its relation to the *regular d-sided polygons*, which is also valid for $\mathcal{S}$. Thus all *triplet* $\mathcal{S}$, $\mathcal{S}_3(q)$, are the *centered hexagonal numbers*, $H_m^c = 1 + 6T_m$ (OEIS A003215) (Fig. 1) [22-24], where $m = \frac{q-1}{3} + 1$ and $T_m = \frac{m(m+1)}{2}$ is the m^{th} triangular number (OEIS A000217), then all *quadruplet* $\mathcal{S}$, $\mathcal{S}_4(q)$, are the *m-fold centered square numbers*, $mQ_m^c = m(1 + 4T_m)$ where $m = \frac{q-1}{2} + 1$ (OEIS A001844), while the *doublet* $\mathcal{S}$, $\mathcal{S}_2(q)$, in principle correspond to Q^c [23].

[3] In our previous papers [22,23], we assumed by the *Self-similar numbers* generalization of decimal number 037 both for a different base of numeral system (and also a digit multiplicity) – the *self-similar analogues* (Fig. 1 in [23]), and for the same numeral system – the *self-similar varieties* (Tab. 3 in [23]). Hereinafter, we assume that the Shcherbak numbers are only referring to the self-similar analogues of 037, while the Self-similar numbers to the self-similar varieties of Shcherbak numbers.

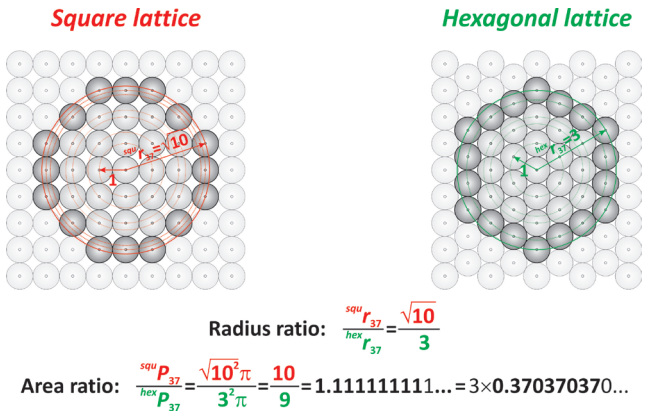


Fig. 1 The theta series of square and hexagonal lattice with respect to lattice point give for quantum 37 the first case of a close-packed circular cluster for both lattices [25]. Area ratio for these two circular clusters is related to the infinite decimal repunit, R_∞, and thus to the Shcherbak number 037.

This relation can be extended to the cyclotomic lattices, vis. N-fold *Bravais lattices*, which in the cases of uniform space filling (the crystallographic cases) correspond to a square lattice ($N = 4$) and an equilateral triangular (hexagonal) lattice ($N = 3{,}6$) [22,23]. In the context of $\mathcal{S}$, Eq. (1), the minimal numeral systems for their realization are $q = 5$ and $q = 4{,}7$, respectively. But if we observe the decimal system as the biquinary system, then $q = 10$ is a minimal system for realization of both symmetries. Indeed, the decimal number 037 is the *closed-pack circular cluster* for both square and hexagonal lattice (Fig. 1).

Beside the generalization of decimal number 037 for the different base of numeral system q and the digit multiplicity n [19,22,23] (Fig. 1 in [23]), we also showed [23] the generalization of $\mathcal{S}$ within a same numeral system as the self-similar varieties of Shcherbak numbers – the *Self-similar numbers* $\tilde{\mathcal{S}}$ (e.g. for decimal number 037 is given in Tab. 1) (fn. 3).

It can be distinguish the two types of $\tilde{\mathcal{S}}$, the r^{th} *vertical self-similar variety of* $\mathcal{S}$,

$$\tilde{\mathcal{S}}_p^{r\uparrow}(q) = \frac{q^{(p-1)r}+q^{(p-2)r}+\cdots+q^r+1}{p} = \frac{R_p(q^r)}{p} = \mathcal{S}_p(q^r),\ r \in \mathbb{N}_{\geq 1} \qquad (3)$$

and the r^{th} *horizontal self-similar variety of* $\mathcal{S}$,

$$\tilde{\mathcal{S}}_p^{r\rightarrow}(q) = \frac{q^{pr-1}+q^{pr-2}+\cdots+q+1}{p} = \frac{R_{p\times r}(q)}{p} = R_r(q)\mathcal{S}_p(q^r).$$

We also showed that $\tilde{\mathcal{S}}^{\uparrow}$, Eq. (3)[4], have a very interesting property of *numerical scaling* (Tab. 1), *geometrical scaling* (e.g. all vertical varieties of the triplet $\mathcal{S}$, $\tilde{\mathcal{S}}_3^{r\uparrow}(q)$, also represent the centered hexagonal numbers; Fig. 3 in [23]) *and arithmetical scaling* (Tabs. 4 and 5 in [23]). For the purposes of this study, the decimal system will be of particular interest, as well its scaling by a power of ten.

Tab. 1. The self-similar varieties of decimal number 037 [23].

Relation between the first three vertical and horizontal decimal varieties of 037		
$\tilde{\mathcal{S}}_3^{r\uparrow}(10)$		$\tilde{\mathcal{S}}_3^{r\rightarrow}(10)$
$\mathbf{037} = \mathbf{111}/3$	$\xrightarrow{\times \mathbf{1}}$	$\mathbf{111}/3 = \mathbf{037}$
$0\mathbf{03}3\mathbf{6}7 = 0\mathbf{10}1\mathbf{0}1/3$	$\xrightarrow{\times \mathbf{11}}$	$111\mathbf{111}/3 = 037\mathbf{037}$
$000\mathbf{3}33\mathbf{6}67 = 00\mathbf{1}00\mathbf{1}00\mathbf{1}/3$	$\xrightarrow{\times \mathbf{111}}$	$111111\mathbf{111}/3 = 037037\mathbf{037}$

Relation between the first three vertical decimal varieties of 037 and centered hexagonal numbers
$\tilde{\mathcal{S}}_3(10) = \mathbf{037} = H_4^c = 6\times T_3 + 1$
$\tilde{\mathcal{S}}_3^{2\uparrow}(10) = 0\mathbf{03}\mathbf{6}7 = H_{34}^c = 6\times T_{33} + 1$
$\tilde{\mathcal{S}}_3^{3\uparrow}(10) = 000\mathbf{3}33\mathbf{6}67 = H_{334}^c = 6\times T_{333} + 1$
$T_m - m^{\text{th}}$ triangular number, $H_m^c - m^{\text{th}}$ centered hexagonal number.

The last particular property of $\mathcal{S}$ to be mentioned in this paper is their relation to the *generalized golden mean*. Namely, the cyclotomic polynomial, Eq. (2), can be regarded as a complementary form of generalized *golden polynomial* [22,23]

$$\Phi_n(q) = q^{n-1} - q^{n-2} - \cdots - 1,$$

whose largest root on the open interval (1, 2) is the generalized golden mean [26]. In the case of the basic form of golden mean,

$$\Phi = \Phi_3(10) = \frac{1+\sqrt{5}}{2} \text{ and } \phi = \phi_3(10) = \frac{-1+\sqrt{5}}{2}, \tag{4}$$

and its scaled golden polynomials,

$$\Phi_3(q^r) = q^{2r} - q^r - 1, \tag{5}$$

$$\bar{\Phi}_3(q^r) = \Phi_3((-q)^r) = q^{2r} + q^r - 1 = \phi_3(q^r), \tag{6}$$

complementary cyclotomic polynomials are obtained,

$$\Phi_3(q^r) = q^{2r} + q^r + 1, \tag{7}$$

$$\bar{\Phi}_3(q^r) = \Phi_3((-q)^r) = q^{2r} - q^r + 1. \tag{8}$$

[4] Eq. (3) refers to the two different numbers, vis. the numbers given in a different notation, with the same value. If we introduce a notation for the higher digits as $\mathrm{A} = \overline{10}$, $\mathrm{B} = \overline{11}$, ..., then for the first non-trivial vertical variety of $\mathcal{S}_3(10) = \mathbf{037}$ holds $\tilde{\mathcal{S}}_3^{2\downarrow}(10) = 0\mathbf{03}\mathbf{6}7 = \mathcal{S}_3(100) = \overline{0\mathbf{0}}\ \overline{3\mathbf{3}}\ \overline{6\mathbf{7}}$.

These two types of complementary polynomials that differ only in the signs, give two complementary solutions. Eqs. (5) and (6) respectively relate to the *golden integers* 89 and 109 with their scaled values, which generate the Fibonacci numbers (Tab. 6 in [23]). Eqs. (7) and (8) respectively relate to the cube roots of unity and *cyclotomic values* 111 and 91 with their scaled values, which generate 037 and its decimal varieties (Tab. 6 in [23]).

All these relations based on a self-similarity and a scaling indicate the need for a deeper insight into the relationships between the decimal number 037 (the quantum 037) and the golden mean (the constant ϕ).

3. The number 037 and the golden mean

The simplest idea of a number from which began the evolution of numbers refers to the measurement process – a *counting*. Counting as a repeated addition by 1 is inherently *dynamical* process and as such is an underlying basic concept of the Peano axiomatization of natural numbers $\mathbb{N}$ through a primitive recursive function, vis. the *successor function* or *successor operation* S such that $S(n) = n + 1$. The axiomatizing $(1, S(1)) \in \mathbb{N}$ defines a unary representation of the counting numbers $\mathbb{N}_{\geq 1}$ since $1_1 = 1$, $11_1 = S(1)$, $111_1 = S(S(1))$,... A further crucial step in the development of numbers consists in introducing of *numeration* concept by which arbitrary large unary numbers can be encoded by finitely many symbols, i.e. digits. This disparity between the general infiniteness of unary numbers and finiteness of digit set "…requires some kind of invariance of the representation and a recursive algorithm which will be iterated, hence something of a dynamical kind again. … Moreover, the representation could be obtained in a purely dynamical way and had a meaning in terms of modular arithmetic" [27]. The modularity, as a form of *circularity*, in a numeration is enabled by two great inventions – a positional notation and the zero. For the classical base-q numeral systems it results in a defining of the *digit set* E,

$$E := \{k \in \mathbb{N}: 0 \leq k \leq q - 1, q \geq 2\},$$

and the q-*adic representation* of non-negative integer $n \in \mathbb{N}$,

$$n := \epsilon_0(n)\epsilon_1(n) \cdots \epsilon_{L-1}(n)\epsilon_L(n) = (\epsilon_j(n))_{0 \leq j \leq L}, \quad \epsilon_j(n) \in E, L \in \mathbb{N},$$

given by the q-*adic expansion* which satisfies both modular and positional principle,

$$n = \epsilon_0(n) + (\epsilon_1(n) + \cdots + (\epsilon_{L-1}(n) + \epsilon_L(n)q)q \cdots)q$$
$$= \sum_{j=0}^{L} \epsilon_j(n)q^j.$$

The q-adic numeration from a dynamical viewpoint is concerned with counting, vis. a transformation of number representation under the successor function, on the compact group of integers and it is given by a concept of the *q-adic odometer* ("adding machine") [27,28]. The compactification of $\mathbb{N}$, using an injective mapping $n \mapsto \tilde{n}$ from $\mathbb{N}$ to the infinite compact product space $E^{\mathbb{N}}$ given by

$$\tilde{n} := \epsilon_0(n)\epsilon_1(n)\cdots\epsilon_{L-1}(n)\epsilon_L(n)0^{\infty} = n0^{\infty} = (\epsilon_j(n))_{j\geq 0} \qquad (9)$$

(the string $\tilde{n}$ ends with an infinite sequence of digit 0 by putting $\epsilon_j(n) = 0$ for all $j > L$), enables to indentify $\mathbb{N}$ with its image $\widetilde{\mathbb{N}}$ which is a dense subset of $E^{\mathbb{N}}$ (in the sequel, we will indentify n with $\tilde{n}$).

Now q-adic odometer can be defined as the dynamical system $(\mathcal{K}_q, \tau)$, where the set of representations $\mathcal{K}_q := \{\epsilon_0\epsilon_1\cdots\epsilon_{L-1}\epsilon_L 0^{\infty} : L \in \mathbb{N}, \epsilon_j \in E\}$ is a compact subspace of $E^{\mathbb{N}}$ homeomorphic to a countable initial segment of the ordinals and the addition-by-one map $\tau: \widetilde{\mathbb{N}} \mapsto \widetilde{\mathbb{N}}$ is given by $\tau\big((\epsilon_j(n))_{j\geq 0}\big) := (\epsilon_j(n+1))_{j\geq 0}$. Since each n on $\mathcal{K}_q$ has a successor given by $n^+ = \tau(n)$, then $\tau^{-1}(0^{\infty})$ is the set of maximal points, which for q-adic case is $\tau^{-1}(0^{\infty}) = \{(q-1)^{\infty}\}$. Thus the digit strings $(q-1)^t$ (the representation of $q^t - 1 = (q-1)R_t(q)$; Eq. (2)) contain important information concerning carry propagation when adding 1 and especially reflect properties of the q-adic odometers.

The compactification, Eq. (9), enables bijective q-adic numeration (an avoiding of multiple representation such as "1", "01", "001", … for the same value 1) and corresponds to the Hensel expansion[5] of n, i.e. an infinite q-adic expansion which represents $\mathbb{Z}_q$, and thus embedding $\mathbb{N}$ in $\mathbb{Z}_q$. Using a geometric description of Hensel's q-adic numbers [29] gives for the decimal system an interesting relation between the golden mean and the number 037 (Fig. 2).

Let us associate to each q-adic integer n, Eq. (9), a vector map $\lambda: \mathbb{Z}_q \to \mathbb{C}$ as

$$n = \sum_{j\geq 0} \epsilon_j(n) q^j \;\longmapsto \lambda(n) = \sum_{j\geq 0} \frac{\mathbf{e}_{\epsilon_j(n)}}{\nu^j}, \qquad (10)$$

where $\mathbf{e}_{\epsilon_j}$ is a digit position vector given by $\mathbf{e}_k = e^{2\pi i k/q}$, $k \in E$, and ν is a scaling factor such that $\nu > 2\Phi + 1$, with Φ defined by Eq. (4), when it enables the mapping λ is injective. Then decimal counting unit would correspond to the *tenth* roots of unit and to a *decagon* with unit radius (Fig. 2A).

Let us now introduce a discretization of the real line such that a mapping from a continuous to discrete domain be unique, i.e. a discretization must be comprised of non-overlapping intervals and to spans each element of the domain. Define a surjective discretization mapping $\Delta: \mathbb{R} \to \mathbb{Z}$

[5] For Hensel's q-adic numbers, an integer q is commonly restricted to the primes, since it is possible for a composite number q to find pairs of non-zero q-adic numbers whose product is 0.

$$\Delta(x) = \{i \in \mathbb{Z}: x = i + r, r \in I = [0,1)\}, \tag{11}$$

where integer i denotes a *position* of the discretization interval $\delta_i = i + I$ (Eq. (11) also represents the one of many definitions of the *floor* (*greatest integer*) *function* $\lfloor x \rfloor$ which value at x is the integer (integral) part of x). Each $r \in I$ is a representative of exactly one equivalence class or a coset of the quotient group $\mathbb{R}/\mathbb{Z}$, so that I is a set of representatives of all the cosets. Thus the result of the operation Δ on the real line is a "projection" of $\mathbb{R}$ on $\mathbb{Z}$, i.e. $\Delta(\mathbb{R}) = \Delta(\bigcup_{i\in\mathbb{Z}} \delta_i) = \mathbb{Z}$, which can be reduced to a "projection" of I on 0. Division by powers of q in the set of representtatives enables denser discretization (the scaling of the discretization),

$$\Delta_\ell(x) = \{i \in \mathbb{Z}: x = \tfrac{i+r}{q^\ell},\ r \in [0,1), q \in \mathbb{N}_{\geq 2}, \ell \in \mathbb{N}\}, \tag{12}$$

so when ℓ tends to an arbitrary large value then Eq. (12) yields an arbitrary "fine" discretization, while it is reduced to Eq. (11) for $\ell = 0$.

Now it is possible to associate the q-adic odometer with the numeration process of discrete real line. Let an association of a q-adic digit $\epsilon \in E$ with the ith discretezation interval $\delta_i \in \mathbb{R}$ denote by $\epsilon^{(i)}$, then we can introduce a mapping function $\varphi: \mathcal{K}_q \longrightarrow \mathbb{Z}$ which sends each digit of a digit sequence of the integer n to the corresponding position on the discrete real line such that

$$\varphi\big((\epsilon_j(n))_{j\geq 0}\big) := (\epsilon_j^{(n-j)}(n))_{j\geq 0}, \tag{13}$$

by which we assume that the negative integers $-n$ differ from the positive integers n just by position on the real line. Superposition of the overlapping sequences for the first N nonnegative integers yields a unique natural number

$$\mathcal{p}_N(q) = \sum_{n=0}^{N} \sum_{j=0}^{n} \epsilon_j(n) q^{N-(n-j)}, \tag{14}$$

which can be also obtained by the recurrence formula $\mathcal{p}_N(q) = q\mathcal{p}_{N-1}(q) + N$ with the initial condition $\mathcal{p}_0(q) = 0$, what with Eq. (2) results

$$\mathcal{p}_N(q) = \sum_{n=0}^{N} (N-n) q^n = \frac{q(q^N-1)-N(q-1)}{(q-1)^2} = \frac{qR_N^2(q)}{q^N-1} - \frac{N}{q-1}. \tag{15}$$

From Eq. (13) follows that the corresponding equation of Eq. (14) for the negative part of discrete real line, i.e. for $i \in \mathbb{Z}_{<0}$, becomes

$$\bar{\mathcal{p}}_N(q) = \sum_{n=1}^{N} n q^{n-1} = \frac{Nq^N(q-1)-(q^N-1)}{(q-1)^2} = \frac{Nq^N}{q-1} - \frac{R_N^2(q)}{q^N-1}, \tag{16}$$

where Eq. (16) is also the result of recurrence relation $\bar{\mathcal{p}}_N(q) = Nq^{N-1} + \bar{\mathcal{p}}_{N-1}(q)$ with the initial condition $\bar{\mathcal{p}}_0(q) = 0$.

From Eqs. (15) and (16) follow two main relations

$$\mathcal{p}_N(q) + q\bar{\mathcal{p}}_N(q) = NR_{N+1}(q) \tag{17}$$

and

$$\begin{aligned} D_N(q) = R_N^2(q) &= \mathcal{p}_N(q)q^{N-1} + \bar{\mathcal{p}}_{N-1}(q) = \mathcal{p}_{N-1}(q)q^N + \bar{\mathcal{p}}_N(q) \\ &= \mathcal{p}_N(q)q^{N-1} - Nq^{N-1} + \bar{\mathcal{p}}_N(q) \\ &= \mathcal{p}_{N-1}(q)q^N + Nq^{N-1} + \bar{\mathcal{p}}_{N-1}(q), \end{aligned} \tag{18}$$

where $D_N(q)$ is the square of repunit $R_N(q)$ and so called the *(wonderful) Demlo number* [30] (OEIS A002477).

Eq. (17) shows that the digits of consecutive and reverse integer sequence $\mathcal{p}_N(q)$ and $q\bar{\mathcal{p}}_N(q)$, respectively, due to a carryover propagation generally are the $\mathrm{dr}(N)$-complements, where the function $\mathrm{dr}(n) = n - 1 (\mathrm{mod}\, q-1) + 1$, assuming that $q - 1 (\mathrm{mod}\, q-1) \equiv 0$, is the *digital root* or *repeated digital sum* (OEIS A010888). The exceptions to this general rule for the ℓ-digit number N relate to the $(\ell-1)$-end digits of $q\bar{\mathcal{p}}_N(q)$ as a consequence of the carries, too.

Furthermore, Eq. (18) shows that $\mathcal{p}_N(q) = q\mathcal{p}_{N-1}(q) + N$ and $\bar{\mathcal{p}}_N(q) = Nq^{N-1} + \bar{\mathcal{p}}_{N-1}(q)$ are the constitutive parts of $D_N(q) = R_N^2(q)$. Namely, the lattice multiplication makes an obvious fact that the square of each repunit $R_N(q)$ is an almost palindrome number (an exact for $N < q$) which consists of the $(2N-1)$-digit string divided into the two parts: the initial $(N-1)$-digit part which is a periodically strictly ascending ordered string, $\mathcal{p}_{N-1}(q)$, and the final $(N-1)$-digit part which is a periodically strictly descending ordered string, $\bar{\mathcal{p}}_{N-1}(q)$, while the central digit which has the weight q^{N-1} is a locally maximal (except when is 0) and hence can be assigned to both parts. More precisely, the central digit can be get from the modified digital root $\widehat{\mathrm{dr}}(N) = n - 2(\mathrm{mod}\, q-1) + 2$ (OEIS A117230)[6], which for the special case $\widehat{\mathrm{dr}}(m(q-1)+1) = 1q + 0 = (1)0$ means that the digit with the weight q^{N-1} equals 0 while the carryover 1 transfers to the digit with the weight q^N, generally implicated that the central digit of $D_N(q)$ can never be 1 (except in the trivial case $R_1(q) = 1$). Moreover, the presented determination of a digit value for $N-1$ number position of $D_N(q)$ is rather a general rule due to a carryover propagation, so that

$$\begin{aligned} D_N(q) &= q^N \textstyle\sum_{j=1}^{N-1} \widehat{\mathrm{dr}}(j) q^{N-j-1} + \widehat{\mathrm{dr}}(N) q^{N-1} + \sum_{j=1}^{N-1} \widehat{\mathrm{dr}}(j) q^{j-1} \\ &= q^N \vec{D}_N(q) + \overleftarrow{D}_N(q), \end{aligned} \tag{19}$$

where $\vec{D}_N(q)$ and $\overleftarrow{D}_N(q)$ respectively denote a numerical representation of the ascending and descending part of $D_N(q)$ with completed carryover propagation.

Obviously, the modified digital root $\widehat{\mathrm{dr}}(j)$ of the consecutive and reverse non-negative integers results in the $(q-1)$-digit periods $\vec{P}(q)$ and $\overleftarrow{P}(q)$, respectively,

[6] By $\widehat{\mathrm{dr}}(N)$, we could still imply $\mathrm{dr}(N)$, but under the assumption $q - 1 (\mathrm{mod}\, q-1) \equiv q - 1$.

which depend only on the base q. From Eqs. (17) and (19) follows that the periods have the form

$$\vec{P}(q) = \wp_{q-2} + 1 = (012 \cdots q-3\ q-1)_q, \tag{20}$$

$$\overleftarrow{P}(q) = \bar{\wp}_{q-1}(q) - 1 = (q-1\ q-2 \cdots 320)_q. \tag{21}$$

and that their mutual dependence can be expressed by

$$\vec{P}(q) + \overleftarrow{P}(q) = q^{q-1} - 1 = (q-1)R_{q-1}(q), \tag{22}$$

$$\overleftarrow{P}(q) = (q-2)q\vec{P}(q), \tag{23}$$

which means that the corresponding digits of the periods $\vec{P}(q)$ and $\overleftarrow{P}(q)$ are the $(q-1)$-complements. In general case, from Eqs. (17) and (19) it can be shown that $\vec{D}_N(q)$ and $\overleftarrow{D}_N(q)$ are $\mathrm{dr}(N)$-complements, i.e.

$$\vec{D}_N(q) + \overleftarrow{D}_N(q) = \mathrm{dr}(N)R_N(q).$$

For the special case of $\vec{D}_M(q)$ and $\overleftarrow{D}_M(q)$ when $q-1|M$, Eqs. (19)-(22) give

$$\vec{D}_M(q) = \left(\vec{P}(q)\right)^{\frac{M}{q-1}} - 1 \quad \text{and}$$

$$\overleftarrow{D}_M(q) = \left(\overleftarrow{P}(q)\right)^{\frac{M}{q-1}} + 1,$$

where their corresponding digits are the $(q-1)$-complements,

$$\vec{D}_M(q) + \overleftarrow{D}_M(q) = q^M - 1 = (q-1)R_M(q).$$

All previous equations in this Section indicate that in each q-adic numeral system exists a unique pure periodical sequence which is the result of a numeration process and it can be considered as the *fundamental sequence (period) of q-adic numeral system*,

$$\mathcal{P}(q) = \lim_{M\to\infty} \vec{D}_M(q) = \left(\vec{P}(q)\right)^{\infty} = (012 \cdots q-3\ q-1)_q^{\infty}. \tag{24}$$

Since the infinite repunit $R_\infty(q)$ can be obtained as a fractional part of the multiplicative inverse of the number $q-1$, i.e.

$$R_\infty(q) = \frac{1}{q-1} \bmod 1, \quad \text{then} \tag{25}$$

$$\mathcal{P}(q) = \frac{1}{(q-1)^2} \bmod 1, \tag{26}$$

where Eqs. (25) and (26) are also related to the generating function of the constant sequence, $G(1;x) = \frac{1}{1-x}$, and the sequence of natural numbers, $G(n;x) = \frac{1}{(1-x)^2}$.

Thereby Eqs. (25) and (26) show a well-known fact that there is a perfect correspondence between the pure periodic integers and the expansion of rational numbers whose denominators are coprime to the base q of numeral system, what is always fulfilled for the successive integers, concretely $q-1 \nmid q$. Hence, the number $q-1$, its divisors and their products have a special role in generating of an invertible residue classes and the cyclic additive groups in a q-adic expansion. In decimal system, this is satisfied for the natural number k when $\gcd(10,k)=1$, i.e. for $k \in \{1,3,7,9\}$, where the cases for $\bar{k}^* \in \{9,7\}$ are complementary to $k^* \in \{1,3\}$ since $\gcd(10,k)=\gcd(10-k,k)$ (Fig. 2A).

If we now by $h \in \mathbb{N}$ and $\Lambda=\{\lambda_i\} \subset \mathbb{N}$ respectively denote a *shift value* and a *set of starting points* for translation $\tau_{h,\Lambda}\left(\{(\epsilon_j(\lambda_i))_{j\geq 0}\}\right) := \{(\epsilon_j(\lambda_i+h))_{j\geq 0}\}$ on $\mathcal{K}_q$ and with $\mathcal{P}_{h,\Lambda}(q)$ a result of superimposing the overlapping sequences $\tau_{h,\Lambda}^{\infty}$, then $\mathcal{P}(q)=\mathcal{P}_{1,\{0\}}(q)$, Eqs. (24) and (26). Again in decimal system, the characteristic periodic sequences are for $h=k^*$ (Fig. 2B), i.e.

$$\mathcal{P}(10)=\mathcal{P}_{1,\{0\}}(10)=\vec{D}_{\infty}(10)=(012345679)^{\infty}, \tag{27}$$

$$\mathcal{P}'(10)=\mathcal{P}_{3,\{0\}}(10)=3\vec{D}_{\infty}(10)=(037)^{\infty}, \tag{28}$$

and the complementary ones for $h=\bar{k}^*$,

$$\mathcal{P}_{7,\{0\}}(10)=7\vec{D}_{\infty}(10)=(086419753)^{\infty}, \tag{29}$$

$$\mathcal{P}_{9,\{0\}}(10)=9\vec{D}_{\infty}(10)=1^{\infty}. \tag{30}$$

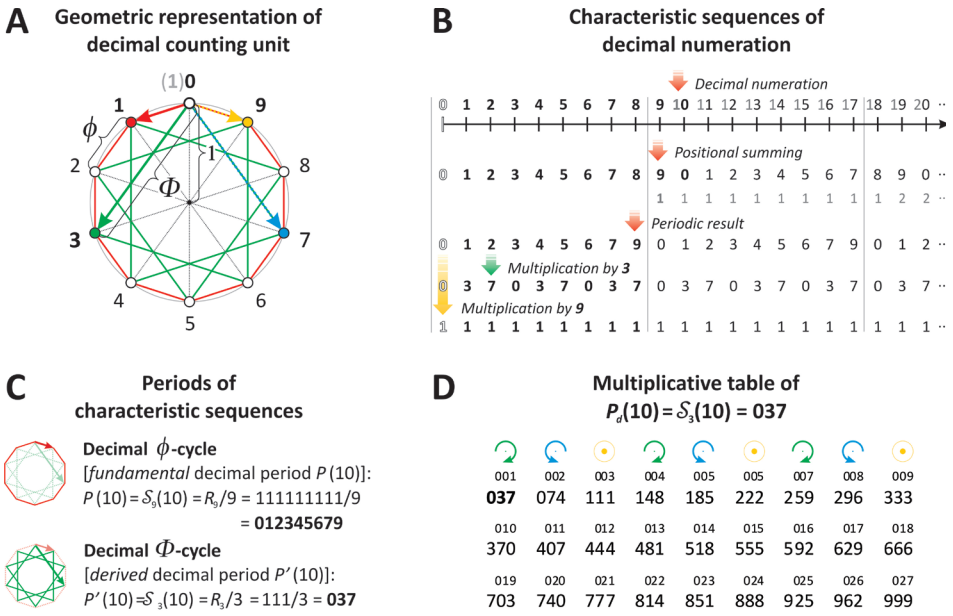


Fig. 2 Geometrical interpretation of decimal numeration process and its characteristic periods (an explanation in the text) [24].

We can now in the decimal system establish a correspondence between the golden mean and the number 037 as follows: in a geometrical representation of the dynamical system $(\mathcal{K}_{10}, \tau)$, the circular shifts by $\mathbf{e}_{k^*} = e^{2\pi i k^*/10}$ for $k^* \in \{1,3\}$, Eq. (10), respectively correspond to the linear shifts by the *golden means* Φ and ϕ, Fig. 2A and Eq. (4), which gives the characteristic periodic sequences $\mathcal{P}(10)$ and $\mathcal{P}'(10)$, Figs. 2B and 2C as well as Eqs. (27) and (28), both determined by the number 037:

$$\vec{P}(10) = \vec{P}_{1,\{0\}}(10) = 012345679 = 037 \cdot 000333667 \quad \text{and} \qquad (31)$$

$$\vec{P}'(10) = \vec{P}_{3,\{0\}}(10) = 037. \qquad (32)$$

Moreover, the fundamental decimal period $\vec{P}(10)$ is a product of the Shcherbak number $\mathcal{S}_3(10) = 037$ and its self-similar variety $\tilde{\mathcal{S}}_3^{3\uparrow}(10) = 000333667$, what is generally fulfilled for the diagonal analogues, vis. when $q = n^2 + 1$ [22,23]. Similarly is valid for $\bar{k}^* \in \{7,9\}$ (Fig. 2A) and corresponding Eqs. (29) and (30):

$$\vec{P}_{7,\{0\}}(10) = 086419753 = 7 \cdot 037 \cdot 000333667 \quad \text{and}$$

$$\vec{P}_{9,\{0\}}(10) = 111 = 3 \cdot 037. \qquad (33)$$

An important fact is that in any q-adic numeral system, due to carryovers, Eq. (28) will be generally satisfied in the form

$$\mathcal{P}_{q-1,\{0\}}(q) = \mathrm{R}_{\infty}(q) = 1^{\infty},$$

which means that in the numeration process is established a *digital complementarity in regards to* $q-1$, in contrast to an *arithmetical complementarity which is in regards to* q. Hence, for the characteristic periods $\vec{P}(10)$ and $\vec{P}'(10)$, Eqs. (31) and (32), their digital complements are respectively

$$\vec{P}_c(10) = \vec{P}_{8,\{0\}}(10) = 098765432 = \overleftarrow{P}(10)/10 \quad \text{and} \tag{34}$$

$$\vec{P}_c'^{(10)} = \vec{P}_{6,\{0\}}(10) = 074 = 2 \cdot 037, \tag{35}$$

where Eq. (34) is the result of previously derived Eq. (23). Generally, multiplication by $q-2$ closely transforms an ascending-descending sequence of $D_N(q)$ into a descending-ascending sequence of its $(q-1)$-complement $D_N^c(q)$, Eq. (22).

A changing of the set of starting points always can be reduced to $\Lambda = \{0\}$ with a corresponding modification of the shift value h, e.g.

$$\mathcal{P}_{3,\{1\}}(10) = \mathcal{P}_{12,\{0\}}(10) = (3 \cdot 4)\vec{D}_{\infty}(10) = (148)^{\infty}, \tag{36}$$

$$\mathcal{P}_{3,\{2,3\}}(10) = \mathcal{P}_{51,\{0\}}(10) = (3 \cdot 17)\vec{D}_{\infty}(10) = (629)^{\infty}. \tag{37}$$

For $3|h$, the periods of periodic parts of the resulting sequences correspond to the multiplicative table of number 037 (Eqs. (32), (35)-(37) and Fig. 2D), while for an arbitrary shift h correspond to a multiplicative table of the number 012345679. If we take into account Eqs. (31), (32) and (33), it follows that *the number 037 is the only number in the decimal system that is contained in each periods resulted from the numeration process, except for the one-digit periods of constant sequences, but it is also true if we take them in a three-digit notation*. This property can be generally applied to the Shcherbak numbers in other q-adic numeral systems taking into account an appropriate number of digits [22-24], what leads to the general conclusion that *the cyclic equivariability as a unique property of the Shcherbak numbers* $\mathcal{S}$ *and the self-similar numbers* $\tilde{\mathcal{S}}$ *has its origin in the sequences which arise from the* ***numeration process***. However, *a unique characteristic of the decimal system is that a digit sum of the fundamental period is also equal to the Shcherbak number*, i.e. $s_{10}(012345679) = 037$.

4. The number 037 and the fine structure constant

Now the question arises of whether we can make a more direct connection between the golden mean and the decimal number of 037. The peculiarity of the golden mean lies in the fact that the difference between the golden mean Φ and its multiplicative inversion ϕ is 1, i.e. $\Phi\phi = 1$ and $\Phi - \phi = 1$ (Fig. 3). As a consequence the sequence of golden mean degrees, $(\Phi^i)_{i\in\mathbb{Z}}$ or its equivalent $(\phi^i)_{i\in\mathbb{Z}}$, has very interesting and unique properties: on the one hand, the sequence is a *geometric progression*, and on the other hand, the sequence has the *property of "additivity"* because each element is equal to the sum of the two preceding elements [31]. If we now determine the constants Ψ and ψ such that they satisfy the previous two properties, but with an additional property which is related to a ***decimal scaling***, i.e. $\Psi\psi = 10$ and $\Psi - \psi = 1$, then we would get the analogy shown in Fig. 3.

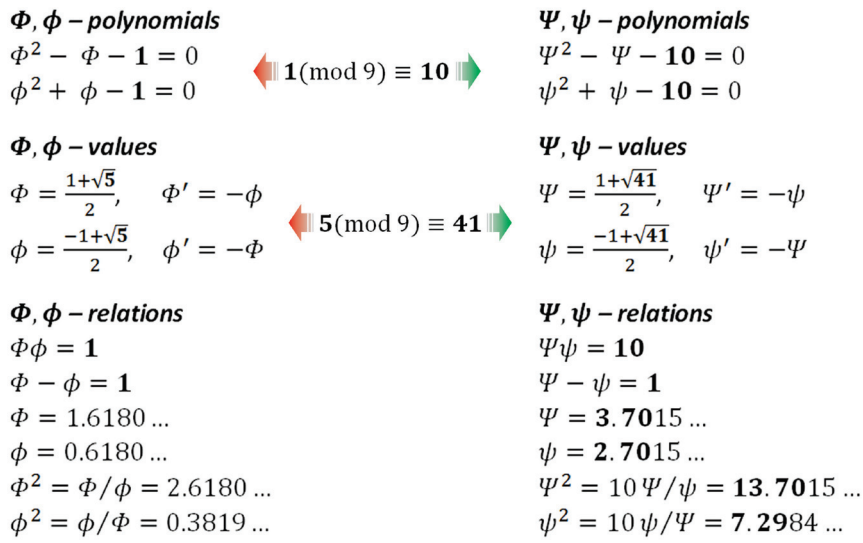


Fig. 3 The analogical properties between the golden means Φ, ϕ and the constants Ψ, ψ.

Beside these basic analogical algebraic properties (Fig. 3), the constants Φ, ϕ and Ψ, ψ share many others related to a problem of the line division, a representation in the form of continued fraction and nested radical, the generalized Fibonacci numbers, the generalized Binet Formulas, etc. Moreover, the constants Ψ and ψ can be regarded as a special case of the *Metallic Means Family* introduced by the quadratic algebraic equation $x^2 - ax - b = 0$ which gives an infinite set

of positive quadratic irrationals for the different values of a and b, as well as generalized the Fibonacci recursive relation $\tilde{F}(n+1) = a\tilde{F}(n) + b\tilde{F}(n-1)$ [32,31], while this special case is realized for $a = \pm 1$ and $b = 10$.

Related to the problem of line division, a definition of Ψ and ψ, given by the equations in Fig. 3, enables a finer division through the *trichotomy principle* instead of the *dichotomy principle* of the golden means Φ and ϕ. Namely, from all possible solutions of the Φ, ϕ-polynomials, the only one belongs to the initial segment, $\phi \in [0,1]$, in contrast to the Ψ, ψ-polynomials which give two solutions, $\Psi, \psi \in [0,10]$. Thus the constants Ψ and ψ enable not only self-similar division, but also an obtaining of middle segment $[\psi, \Psi]$ that is ***ten*** times smaller than the initial segment $[0,10]$. If we consider a "dynamic" model of the trichotomy principle in a form of infinite division or multiplication of the initial segment, then we can regard the middle segment $[\psi, \Psi]$ as the initial segment $[0,1]$ of a "*finer level*" and similarly the initial segment $[0,10]$ as the middle segment $[10\psi, 10\Psi]$ of a "*coarser level*". This recurrent scaling $[\psi, \Psi] \to [0,1]$ and $[0,10] \to [10\psi, 10\Psi]$ can be described by

$$\Psi_{(\ell)}^2 - 10^\ell \Psi_{(\ell)} - 10^{2\ell+1} = 0, \tag{38}$$

$$\psi_{(\ell)}^2 + 10^\ell \psi_{(\ell)} - 10^{2\ell+1} = 0, \quad \ell \in \mathbb{Z}. \tag{39}$$

For $\ell = 0$, Eqs. (38) and (39) are reduced to basic Ψ, ψ-polynomials and values (Fig. 3), while for some arbitrary integer ℓ will be satisfied

$$\frac{\Psi_{(\ell)}}{\psi_{(\ell)}} = \frac{10^\ell \Psi_{(0)}}{10^\ell \psi_{(0)}} = \frac{\Psi}{\psi} = \frac{\Psi^2}{10} = 1.37015 \ldots \quad \text{and} \tag{40}$$

$$\frac{\Psi_{(\ell+1)}}{\psi_{(\ell-1)}} = \frac{10\Psi_{(\ell)}}{\psi_{(\ell)}/10} = 10\Psi^2 = 137.01562 \ldots \tag{41}$$

If we replace the numerical values in Eq. (41) using Eq. (40), then follows

$$\frac{\Psi_{(\ell+1)}}{\psi_{(\ell-1)}} = \frac{\Psi_{(1)}}{\psi_{(-1)}} = \frac{10\Psi_{(0)}}{\psi_{(0)}/10} = \frac{37{,}0156\ldots}{0{,}27015\ldots} = 37.01562 \ldots \cdot 3.70156 \ldots$$
$$= 10\Psi^2 = \Psi_{(1)}^2/10 = 137.01562 \ldots, \tag{42}$$

what the connection between the ***integers*** $\mathcal{S}_3(10) = \mathcal{S} = 037$ and $\mathcal{s} = 027$, on the one side, and the ***irrationals*** $\Psi_{(1)} = 37.015 \ldots$ and $\psi_{(1)} = 27.015 \ldots$, on the other side, makes an indicative. Their corresponding quadratic algebraic equations,

$$\Psi_{(1)}^2 - 10\Psi_{(1)} - 10^3 = 0 \quad \text{and} \tag{43}$$

$$\mathcal{S}^2 - 10\mathcal{S} - (10^3 - 1) = 0, \tag{44}$$

as well as the complementary ones,

$$\psi_{(1)}^2 + 10\psi_{(1)} - 10^3 = 0 \quad \text{and} \tag{45}$$

$$\mathcal{s}^2 + 10\mathcal{s} - (10^3 - 1) = 0, \tag{46}$$

reveal that Eqs. (44) and (46) are the minimal (***unit***) variation of Eqs. (43) and (45) for which it is possible to obtain an integer solution, and thus $\mathcal{S}$ and $\mathcal{s}$ are the smallest possible integers which satisfy the self-similarity and the scaling by powers of 10. The importance of previous is in a fact that generally *the irrationals are related to the continuous quantities, while the integers to the discontinuous quantities, implicated that this process of transferring continuous equations and models into discrete counterparts is related to* ***discretization***. From this point of view, it is interesting an analogous relation of Eq. (42) for $\mathcal{S}$ and $\mathcal{s}$,

$$\Xi_\alpha = \frac{10\mathcal{S}}{\mathcal{s}/10} = \frac{1}{10}37 \cdot 37.(037)^\infty = 137.(037)^\infty, \tag{47}$$

since its value is very close to a fundamental dimensionless physical quantity, known as the *fine structure constant*, which a standard value[7] can be reduced to $\alpha^{-1} = 137.036$ with the absolute error less than 10^{-6}. It is worth noting that a relative difference of the value Ξ_α^{-1} given by Eq. (47) from the standard value of fine structure constant α_0 is $\frac{\Xi_\alpha^{-1} - \alpha_0}{\alpha_0} = -7.57 \cdot 10^{-6}$ and thus it is significantly less than any well known theoretical and experimental constraints on the fine structure constant related to the time variations, $\frac{\Delta\alpha}{\alpha} = (3.6 \pm 3.7) \cdot 10^{-3}$, and the spatial variations, $\frac{\delta\alpha}{\alpha} = (-2.4 \pm 3.7) \cdot 10^{-2}$ [34]. Also since the strength of the electromagnetic interaction varies with the strength of the energy field, at interaction energies above $90\ GeV$ the fine structure constant is known to approach $1/127$, where this value can be approximate with the relative difference $2.92 \cdot 10^{-4}$ by a similar equation to Eq. (47):

$$\xi_\alpha = \frac{10\mathcal{S} - \mathcal{s}}{\mathcal{s}/10} = \Xi_\alpha - 10 = 127.(037)^\infty, \tag{48}$$

Furthemore, the absolute difference

$$\Xi_\alpha - \alpha^{-1} = \frac{\mathcal{s}+1}{\mathcal{s}}10^{-3} = 1.(037)^\infty \cdot 10^{-3} \tag{49}$$

can be also determined by $\mathcal{s}$ and the characteristic period of decimal numeration $\vec{P}'(10)$, Eq. (28), similarly as in the case of Eqs. (47) and (48). For more serious consideration of this approach, it is it is necessary to find in a much larger extent the relationships, like Eq. (47), and the fine tunings, like Eq. (49).[8,9]

---

[7] The most precise value of α obtained experimentally is $\alpha_0^{-1} = 137{,}035999173(35)$ [33].

[8] One of the nondimensionless fundamental physical constants is the speed of light in vacuum and its approximate value is $c = 299800000\ \frac{m}{s}$ with a relative error $2{,}52 \cdot 10^{-5}$ (an exact value of is $c_0 = 299792458\ \frac{m}{s}$). Using $\mathcal{S}$ and the fundamental decimal period $\vec{P}'(10)$, Eq. (33), we can obtain

5. Discussion

Hypothetical objective validity and interpretation of Eqs. (47)-(49) is meaningful only if we assume that the concept of *self-referential mathematical models* can be the background or underlying models which serve as a basic frame of reference for the information systems and thus, according to the emergent property principle [1-16], also for physical and biological systems. In this sense, the self-referential mathematical models are primarily related to the ideal forms arising from various symmetries and hence principally reflect a static aspect of some real system, and which need to be modified for a kinematic and a dynamic part with a consequent emergence of the symmetry breaking. Eqs. (47) and (48) can be principally and roughly interpret in this manner, Eq. (47) reflects the static part, while Eq. (48) and a difference between the approximate value $\alpha^{-1} = 137.036$ and its standard value α_0^{-1} (fn. 7) reflect a kinematic and a dynamic part (cf. Fig. 1). An importance of this hypothetical correspondence would be that both Eqs. (47) and (48) are "quantized" according the same mathematical model represented by Eqs. (44) and (46). Furthermore, since the constants Ψ and ψ correlate with the uncountable quantities such as space and time measures, while the quanta $\mathcal{S}$ and $\mathcal{s}$ with the countable quantities such as a mass expressed as the number of particles (e.g. a nucleon number for the ordinary matter), as well as that these mathematical quantities are defined by the similar Eqs. (44)-(46), follow that *length, time and mass as the three most fundamental physical quantities can be "quantized" using the same composite symmetry which integrates the self-similarity and the scaling by powers of 10*. If we consider the angular measures expressed in degrees or the circumference of a regular hexagon of side length 10, we can again establish an approximate relationship of ϕ with $\mathcal{S}$ or Ψ:

$$60 \cdot \phi = 37{,}0820\ldots \approx 10\Psi \approx \mathcal{S} = 37 \;\Rightarrow\; \phi \approx \frac{\Psi}{6},$$

where we can also notice a mathematical coincidence that 37 is a centered hexagonal number (Fig. 1). The golden angle which plays a significant role in the theory of phyllotaxis and is defined as the angle that divides a full angle in a golden ratio,

---

$$\Xi_c = \frac{\mathcal{S}}{\vec{P}'(10)} \cdot 10^{14} = \frac{10^{14}}{\tilde{\mathcal{S}}^{3\uparrow}} = \frac{10^{14}}{333667} = 299700000,(299700000)^{\infty},$$

and then absolute difference is

$$\Xi_c - c = 99999{,}7(002999997)^{\infty},$$

where $1/2999997 = 333333666667 \cdot 10^{-18} = \tilde{\mathcal{S}}^{6\uparrow}$ and what are to some extent the analogous results to those of Eqs. (47) and (49), cf. [23].

[9] Planck length $\ell_{\mathrm{P}} = 1.616\,199(97) \cdot 10^{-35}\ m$, as it is known, has approximate value $\Phi \cdot 10^{-35}$, where the scaling by powers of 10 for the golden means is obtained from Eqs. (38) and (39) when the free terms are changed by $10^{2\ell}$.

$$\Phi = \frac{\phi \cdot 360^\circ}{\phi^2 \cdot 360^\circ} = \frac{222{,}492\ldots^\circ}{137{,}507\ldots^\circ} \approx \frac{6\mathcal{S}}{\Psi\mathcal{S}} \Rightarrow \Phi \approx \frac{6}{\Psi},$$

also can be approximately interpret by $\mathcal{S}$ and Ψ, what with the previous relations leading to conclusion that *both linear and angular quantities could be quantified by the multiple mutually dependent invariants Φ, $\mathcal{S}$ and Ψ, as well as their coupled values ϕ, $\mathcal{s}$ and ψ*. Above discussion and its conclusions are still preliminary and requires more specific reconsideration, and therefore we will only point out few facts and the theories supporting them.

Scaling laws (power laws) are omnipresence in the natural phenomena, so their identification and clarification of the mechanisms generate them is an important topic of research in many fields of science. A common mechanism which underlies scaling laws in biological systems, especially allometric scaling laws, seems to be related to self-similarity in their fractal structures and dynamics [35]. In physics, the emergence of power law distributions of certain quantities is particularly pronounced for phase transitions. However, the scaling laws defined by powers of 10 is not a common and the one among the best known is related to the *Dirac-Eddington large-number coincidence* that involves pure numbers of the order of 10^{40} generated naturally from the fundamental parameters of the Universe. It has been recently revealed a new large-number coincidence problem as an ensemble of pure numbers of the order of 10^{122} generated from the same set of fundamental parameters, where the similarities between these two large-number coincidences leading to the hypothetical conclusion on their mutual physical scaling, as well as the existence of scaling law between the mass of the nucleon and the cosmological constant (the vacuum energy density), which was also previously suggested from the considerations of field theory and holographic principles [36]. The consistency of these scaling relations [36] is partly related to a dimensionless physical constant the *proton-to-electron mass ratio*, $\mu = \frac{m_p}{m_e} = 1836.15267245$, whose possible variation would lead to the changes in the strength of the strong force. Since the arithmetical regularities in the genetic code are consistent with a nucleon number (Section 2), we will reformulate this constant as

$$\hat{\beta} = \frac{m_p + m_{n^0}}{m_e} = \frac{2m_n}{m_e} = 3674.83633\ldots = \frac{1}{2{,}72120\ldots \cdot 10^{-4}},$$

where m_{n^0} and m_n are respectively the masses of the neutron and the nucleon, so that with the relative difference of $(7.35 \pm 0.5) \cdot 10^{-3}$ is valid

$$\hat{\beta} \approx 10^2\mathcal{S} \approx 10^4\Psi \approx 10^5\mathcal{s}^{-1},$$

what potentially provides an additional theoretical argument for the scaling by powers of 10.

In the regard of geometrical properties of the decimal number 037 and the hypothesis on *its origin as the nucleon packaging quantum of genetic code constituents in a spacetime geometry*, the most convincing preliminary arguments give a theory of quantum gravity called *causal dynamical triangulations* [37]. The theory relies on the Euclidean quantum gravity and the quantum spacetime representation as a simplicial complex which elementary building blocks are a four-dimensional generalization of triangle, known as 4-simplex or tetrahedral hyperpyramid, where these elements model a region of spacetime on the order of the bare Planck volume. A dynamical character of such quantum spacetime is modeled by an assigning to each simplex an arrow of time and their gluing to causal rules where two simplices must be glued together to keep their arrows pointing in the same direction. In the context of our considerations, this study has led to next significant results: the model needed to include from the outset only the cosmological constant, the spectral dimension of spacetime shades from four (on large scales) to two (on small scales), and spacetime breaks up from a smooth continuum into a gnarled fractal [37]. Importance of the first mentioned result follows from the need to exist a scaling law between the mass of the nucleon and the cosmological constant according to the large-number coincidences [36], field theory and holographic principles, what emphasizes the role of the scaling laws related to a nucleon number and their potential selection role in the complex systems, particularly such as biological systems.

The second result may indicate that the quantum spacetime at small scales dominantly acts as a layered triangular lattice or its dual hexagonal lattice, similarly to a graphite which is made out of stacks of graphene layers that are weakly coupled by the van der Waals forces and thus it can be regarded at small scale as a multilayer graphene. This analogy becomes more significant if we bear in mind that graphene brings together issues in quantum gravity and particle physics, and also from soft and hard condensed matter [38], as well as that structurally similar layered hexagonal lattice exists in a recently discovered the liquid crystalline water which plays crucial role in biological systems, even in the origin of life [39-41]. Futhermore, an analogy of this form of carbon and water with a hypothetical quantum spacetime could be nontrivial in their selection as the fundamental foundations of life, as well as in the consequent selection of genetic code constituents, and a result of the same background mechanism which gradually shapes the coherent Universe.

Finally, the third result related to a smooth transition from large scale homogeneity to small scale fractality can be directly connected with the properties of 37 as a figurate number. The number 37 is centered hexa-gonal number, H_4^c (Fig. 1), and thus a packing cluster of tiles or vertices for the pair of dual regular

tilings – the *triangular* and *hexagonal* tiling, but also some of their superposed forms, e.g. the rhombille tiling. The number 37 is also truncated square number, Q_3^t (Fig. 1) (OEIS A005892) and nearly a packing cluster of self-dual regular tiling – the *square* tiling, but also of some semiregular tilings like the tetrakis square tiling. Further, the number 37 is a centered dodecagonal number and also a star number, $S_3 = H_3^c + 6T_2 = 2H_3^c - 1$ (OEIS A003154) [23], and with its own convex part, H_3^c, corresponds to the trihexagonal tiling. Similarly, it is a convex part of its star number successor, $S_4 = 2H_4^c - 1 = 37 + 36 = 73$, and again jointly correspond to the same tiling. Moreover, $10\psi^2 \approx \mathit{s}^2/10 \approx 73$, which means that squares of s and ψ, together with correlated scaling by 10^n, have a geometrical equivalent! The last peculiarity of the number 37 which we refer, but by which we do not exhaust its entire possible geometrical features, is its ability to generate a *fractal-like star* or a *Koch-like snowflake*, $S_m^{\mathrm{Koch}} = 7S_m + 6H_m^c = 10S_m + 3 = 20H_m^c - 3$, what again corresponds to the trihexagonal tiling. Hence, it is generally valid $S_m^{\mathrm{Koch}}/S_m \to 10$ as well as $S_m^{\mathrm{Koch}}/H_m^c \to 20$, what makes both possible – the scaling by powers of 10 and a halving! A unique property for $S_3^{\mathrm{Koch}} = 373$ and $S_4^{\mathrm{Koch}} = 733$ is that also valid $73/37 \approx \psi^3/10$ and $37 \cdot 73 \approx 10^3\psi$. Thus we can conclude that the convex geometrical equivalents of number 37 enable the uniform plane tilings and their derived space counterparts, while its concave geometrical equivalents can be more associated with the fractal tilings. Physical realization of the presented geometrical properties of the number 37 is the most obvious for the various forms of water, particularly those which are related to mesoscopic structures and dynamics, what once again argue in favour of the central role of water in the origin of life. The hydrodynamic quantum analogs [42] indicate that the mechanism of arising these geometrical equivalents of number 037 in a form of the symmetric discrete structures could be at different size scales related to the emergent phenomena, i.e. the self-sustained coherent patterns of collective behaviour arisen as a result of the dynamical processes.

As a general conclusion, we can point out that an importance of the given mathematical interpretation of the fine structure constant is not so much in its accuracy as its simplicity and clarity of the geometric meaning, whence together with other results it follows that one of the dominant symmetry in the Universe is a composite symmetry which integrates the self-similarity and the scaling by powers of 10, further giving a foundation of the natural information processing systems based on the nested codes and the geometrically based "computation" with a further consequence of fractal structural/dynamical organization and the scaling laws, which presence is recognized both in physical and biological realm.

Acknowledgements

This research has been partially funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, through Projects TR-32040 and TR-35023. The author would like to thank Professor Branko Dragovich for fruitful discussions on the various aspects of physics.

References

[1] G.M. D'Ariano, *Adv.Sci.Lett.* **17** (2012) 130.
[2] D. Oriti (ed.), *Approaches to Quantum Gravity: Toward a New Understanding of Space, Time and Matter*, Cambridge University Press, 2009; L. Smolin, [hep-th/0303185].
[3] G. 't Hooft, in *Salamfestschrift: a collection of talks*, A. Ali, J. Ellis and S. Randjbar-Daemi (eds.), World Scientific, 1993, 284-296; L. Susskind, *J.Math Phys.* **36** (1995) 6377.
[4] J. Maldacena, *Adv.Theor.Math.Phys.* **2** (1998) 231; E. Witten, *Adv.Theor.Math.Phys.* **2** (1998) 253; R. Bousso, *Rev.Mod.Phys.* **74** (2002) 825.
[5] E.P. Verlinde, *JHEP* **04** (2011) 029; T. Padmanabhan, *Class.Quant.Grav.* **21** (2004) 4485; S. Lloyd, [quant-ph/0501135].
[6] G.M. D'Ariano and A. Tosini, *Stud.Hist.Phil.Mod.Phys.* **44** (2013) 294.
[7] G. Chiribella, G.M. D'Ariano and P. Perinotti, *Phys.Rev. A* **84** (2011) 012311; G.M. D'Ariano, *Phys.Lett. A* **376** (2012) 697.
[8] J.A. Wheeler, in *Complexity, Entropy and the Physics of Information*, W. Zurek, (ed.), Addison-Wesley, 1990, 3-28; J.D. Bekenstein, *Sci.Am.* **289** (2003) 58.
[9] G.H. Lewes, *Problems of life and mind*, Truebner, 1874–1879; P.A. Corning, *Complexity* **7** (2002) 18.
[10] Bedeau, M. *Phil.Perspect.* **11** (1997) 375.
[11] F.A. Popp, Q. Gu and K.H. Li, *Mod.Phys.Lett. B* **8** (1994) 1269; J.J. Chang, J. Fisch and F.A. Popp, (eds.), *Biophotons*, Kluwer, 1998; L.V. Beloussov, V.L. Voeikov and V.S. Martynyuk (eds.), *Biophotonics and Coherent Systems in Biology*, Springer, 2007.
[12] D. Abbott, P.C.W. Davies and A. Pati (eds.), *Quantum Aspects of Life*, Imperial College Press, 2008; N. Lambert et al., *Nature Phys.* **9** (2013) 10.
[13] S. Hameroff and R. Penrose, *Phys.Life Rev.* **11** (2014) 39.
[14] M. Tegmark, 2014, arxiv:1401.1219 [quant-ph].
[15] B. Barbieri (ed.), *Introduction to Biosemiotics*, Springer, 2007; *The Codes of Life*, Springer, 2008.
[16] E.N. Trifonov, *Bull.Math.Biol.* **51**(1989) 417; E.N. Trifonov, in *The Codes of Life*, B. Barbieri (ed.), Springer, 2008, 3-14; G. Battail, in *Introduction to Biosemiotics*, B. Barbieri (ed.), Springer, 2007, 299-345.
[17] C.R. Woese, *Proc.Natl.Acad.Sci.USA* **54** (1965) 1546; *Proc.Natl.Acad.Sci.USA* **99** (2002) 8742.
[18] V.I. Shcherbak, *J.Theor.Biol.* **166** (1994) 475.
[19] V.I. shCherbak, *Biosystems* **70** (2003) 187; In *The Codes of Life*, M. Barbieri (ed.), Springer, 2008; 153-188; V.I. shCherbak and M.A. Makukov, *Icarus* **224** (2013) 228.

[20] A.B. Verkhovod, *J.Theor.Biol.* **170** (1994) 327; A.M. Downes and B.J. Richardson, *J.Mol.Evol.* **55** (2002) 476; T. Négadi, *Neuroquantology* **7** (2009) 181.
[21] M.M. Rakočević, *Biosystems* **46** (1998) 283; *J.Theor.Biol.* **229** (2004) 221.
[22] N.Ž. Mišić, *Proc.NEUREL* (2010) 97; *Proc.DOGS* (2010) 132.
[23] N.Ž. Mišić, *Neuroquantology* **9** (2011) 702.
[24] N.Ž. Mišić, *Proc.SAUM* (2004) 122.
[25] N.J.A. Sloane and B.K. Teo, *J.Chem.Phys.* **83** (1984) 6520.
[26] E.P. Miles, *Am.Math.Mon.* **67** (1960) 745.
[27] G. Barat, V. Berthé, P. Liardet and J. Thuswaldner, *Ann.Inst.Fourier* **56** (2006) 1987.
[28] P. Grabner, P.L. and R. Tichy, *Acta Arith*. **70** (1995) 103.
[29] A.M. Robert, *A Course in p-adic Analysis*, Springer-Verlag, 2000.
[30] D.R. Kaprekar, *Math. Student* **6** (1938) 68.
[31] A.P. Stakhov, *Comput.Math.Appl.* **17** (1989) 613; *The Mathematics of Harmony*, World Scientific Publishing, 2009.
[32] V.W. de Spinadel, *From the Golden Mean to Chaos*, Nueva Libreria, 1998; *Visual Mathematics* **1** (1999).
[33] R. Bouchendira et al., *Phys.Rev.Lett.* **106** (2011) 080801.
[34] P.A.R. Ade et al., *Astron.Astrophys.* **561** (2014) A97.
[35] G.B. West, J.H. Brown and B.J. Enquist, *Science* **276** (1997)122.
[36] S. Funkhouser, Proc.R.Soc.A **464** (2008) 1345; S. Capozziello and S. Funkhouser, *Mod. Phys.Lett. A* **24** (2009) 1121; *Mod.Phys.Lett. A* **24** (2009) 1743.
[37] J. Ambjørn, J. Jurkiewicz and R. Loll, *Sci.Am.* **299** (2008) 42.
[38] A.H. Castro Neto et al., *Rev.Mod.Phys.* **81** (2009) 109.
[39] Ling, G.N. (2003) *Physiol.Chem.Phys.Med NMR* **35** (2003) 91.
[40] G.H. Pollack, X. Figueroa, and Q. Zhao, *Int.J.Mol.Sci.* **10** (2009) 1419; G.H. Pollack, *The Fourth Phase of Water: Beyond Solid, Liquid, and Vapor*, Ebner and Sons, 2013.
[41] M. Chaplin, *Nat.Rev.Mol.CellBiol.* **7** (2006) 861.
[42] S. Douady and Y. Couder, *Phys.Rev.Lett.* **68** (1992) 2098; A. Eddi, A. Decelle, E. Fort and Y. Couder, *Europhys.Lett.* **87** (2009) 56002; Y. Couder, A. Boudaoud, S. Protière and E. Fort, *Europhys.News* **41** (2010) 14. G. Pucci, M. Ben Amar and Y. Couder, *J.Fluid Mech.* **725** (2013) 402.

Correlation of T-cell Epitope Location and Order/Disorder Protein Structure

Nenad S. Mitić [a]
Faculty of Mathematics, Studentski trg 16/IV, Belgrade, Serbia

Mirjana D. Pavlović [b]
Institute of General and Physical Chemistry, Studentski trg 12/V, Belgrade, Serbia

Davorka R. Jandrlić [c]
Faculty of Mechanical Engineering, Kraljice Marije 16, Belgrade, Serbia

Saša N. Malkov [d]
Faculty of Mathematics, Studentski trg 16/IV, Belgrade, Serbia

Abstract

Disordered protein regions are extremely sensitive to proteolysis in vitro, and are expected to be under-represented as T-cell epitopes. Since these regions are also prevalently hydrophilic, the

[a] e-mail address: nenad@matf.bg.ac.rs
[b] e-mail address: mpavlovic@iofh.bg.ac.rs
[c] e-mail address: djandrlic@mas.bg.ac.rs
[d] e-mail address: smalkov@matf.bg.ac.rs

aim of our research was to find out whether disorder and hydropathy prediction methods could help in understanding processing and selection of immunodominant epitopes. Here, we focus on the characteristics of epitopes in the consensus of 9 publicly available disorder predictors. Epitopes were predicted by the pan-specific T-cell epitope predictors NetMHCpan and NetMHCIIpan. Frequency of epitopes presented by human leukocyte antigens (HLA) class-I or -II was found to be almost 10 times higher in consensus of ordered than in consensus of disordered protein regions. Both HLA class-I and HLA class-II binding epitopes are prevalently hydrophilic in disordered and prevalently hydrophobic in ordered protein regions, whereas epitopes recognized by HLA class-II alleles are more hydrophobic than those recognized by HLA class-I. As regards both classes of HLA molecules, high-affinity binding epitopes display more hydrophobicity than low affinity-binding epitopes (in both ordered and disordered regions). Epitopes belonging to disordered protein regions were not predicted to have poor affinity to HLA class-II molecules, as expected from disorder intrinsic proteolytic instability. The relation of epitope hydrophobicity and order/disorder location was also valid if alleles were grouped according to the HLA class-I and HLA class-II supertypes, These data suggest that reverse vaccinology, oriented towards high-affinity epitopes, is also oriented towards prevalently hydrophobic epitopes encompassing the consensus of ordered regions. The analysis of predicted and experimentally evaluated epitopes of cancer-testis antigen MAGE-A3 has confirmed that the majority of T-cell epitopes, particularly those that are promiscuous or naturally processed, was located in ordered and disorder/order boundary protein regions overlapping hydrophobic regions.

1 Introduction

The possibility to use peptides or protein antigens (Ags), as the constituents of vaccines, presents many advantages. Peptides and proteins can be easily produced in vitro reducing production costs and simplifying large-scale vaccine production procedures. Moreover, expression of peptides belonging to pathogens overcame pathogens culturing issues. Binding of peptides to MHC molecules, is the most selective step in defining T cell epitopes, and is used for epitope screening. Computational epitope-prediction pro-

grams are trained on the known peptide-binding affinities to a particular MHC molecule . They can be categorized as sequence-based and structure-based methods ([1], [2], [3], [4], [5], [6]). The first category focuses on the primary structure of analyzed Ags and identification of binding peptides, while the second makes use of 3D structure of MHC molecules. Sequence-based methods for the prediction of MHC-binding peptides include binding motifs, quantitative matrices, artificial neural networks, hidden Markov models, and molecular modeling ([1]). Structure-based methods include docking of peptides, threading algorithms, binding energy and molecular dynamics to discriminate between binding and non-binding peptides ([7]). In the course of this study, we have used NetMHCpan and NetMHCIIpan servers ([8], [9]) for the prediction of MHC class-I and MHC class-II T cell epitopes, respectively. These are mixed sequence- and structure-based methods, which contain information from both binding peptides and MHC-binding sites (peptide-contacting polymorphic MHC residues, i.e. MHC "pseudo-sequences").

Immune response does not recognize all possible epitopes in a protein Ag, predicted as MHC binding peptides, but are focused on a relatively small number of epitopes. These epitopes are termed immunodominant. Epitope immunodominance is influenced by the efficiency with which the epitopes are generated by cellular processing, transported and presented on the surface of the Ag-presenting cells (APCs) ([10]). Clusters of immunodominant CD4+ T-cell epitopes were found in limited regions, often within sites adjacent to protease sensitive flexible loops ([11],[12],[13],[14]), or within structurally stable regions ([15], [16]). These data suggest that proteolytic release of peptides may be a key factor determining the epitope density with class-II molecules on the cell surface, influenced by the events known as epitope "context", such as the three dimensional (3D) structure of the Ag, the location of the epitope within the Ag and endoprotease cleavage sites in the residues flanking the epitope ([17], [18]).

In contrast to the studies in favor of the epitope "context", Weaver and colleagues ([19], [20]) have demonstrated that CD4+ T-cell immunodominance was not dependent on Ag 3D structure, but on the intrinsic epitope characteristics and persistence of epitope-MHC class-II complexes on the surface of APC and epitope "editing" within APC. Epitope "editing", (binding of protein fragments with MHC class-II molecules with the help of the protein chaperone, strongly favors the presentation of high stability complexes over those with lower stability ([19], [21]). MHC class-I epitope processing is not expected to be guided by the Ag 3D structure, as proteasomal processing is known to have ATP-dependent protein-unfolding

activity.

Intrinsically disordered/unstructured proteins (IDPs) and protein regions (IDRs) (reviewed in: [39].) lack stable 3D conformation under physiological conditions in isolation ([22]). They were characterized by different experimental techniques, mainly on missing electron density in a solved X-ray crystallography structure and NMR spectroscopy. Disordered regions [1] are heterogeneous from the structural and functional point of view. They can correspond to entropic chains, flexible linkers between folded domains or flexible N and C protein termini or are involved in transcription and cell-signaling ([23]; [24]). IDRs are inherently sensitive to proteolysis without the need for prior denaturation and, hence, supposed to be under-represented as T cell epitopes due to overproteolysis ([25]). However, protein degradation is tightly regulated in vivo by sequestration of proteases in separate compartments (subcellular organelles) and by controlled substrate delivery. No strong correlation between intrinsic protein disorder and shorter protein half-lives of self-proteins has been reported ([26]). The probability of particular protein segment being cleaved by endosomal proteases can be estimated using structural parameters that indicate conformational flexibility, such as elevated crystallographic B-factors, solvent-accessible surface area or hydrogen-deuterium exchange (HX) for backbone amide groups ([15]; [16]). Although MHC class-I epitope processing is not expected to be guided by the Ag 3D structure, IDRs are capable of destabilizing proteins and targeting them to proteasomes without the need for the previous unfolding ([26]). In addition, proteasome proteases cleavage determines that the clusters of MHC class-I epitopes are located in the hydrophobic protein regions. Anchor positions of the nonamer epitopes binding to many HLA class-I supertypes are prevalently hydrophobic ([27]), as well as the most frequent amino acids (AA) on the majority of other nonamer positions ([28]). The N-terminal positions of the HLA class-II binding peptides almost invariably possess one of the seven hydrophobic AA (F, I, L, M, V, W, Y) and both predicted and experimentally found epitopes have medium to high content of hydrophobic residues ([28]). The AA content of ordered protein regions are also enriched in bulky hydrophobic AA ([29]; [24]). Evidently, both class-I and -II T-cell epitopes

[1] Abbreviations used in the paper: O - (consensus) ordered regions; D - (consensus) disordered region; N - neighbouring region to disorder/order consensus region but not itself consensus of ordered/disordered regions (mixture of disorder/order region for various predictors); KD scale - Kyte-Doolittle amino acid hydrophobicity scale; HW scale - Hopp-Woods amino acid hydrophobicity scale; SB - Strong binding; WB - Weak binding; AvgH - Average hydropathy value method; MAA - Majority of amino acids hydropathy method.

are connected with the order-promoting hydrophobic AA.

Disordered regions can be determined by different algorithms classified according to the principle of their operation as: those based on physicochemical properties of AA in proteins as programs FoldUnfold, PONDR (VL-XT, VL3, VSL2, VSL2b), OnD-CRF GlobPlot, DisEMBL, PreLINK, IUPred, FoldIndex, and those based on alignments of homologous protein sequences (RONN, DISOPRED) (reviewed in: [41]; [30]). The majority of these predictors use standard machine-learning techniques. Other approaches implement physical principles governing the process of protein folding. There is no universal solution for comparing them and establishing the "best" predictor. Thus, it is recommended to compare predictions by different algorithms based on different physical and/or computational principles and seek a consensus of their scores, enabled in metapredictors. Metapredictors either simply help carry out numerous parallel predictions of order/disorder, secondary structure and hydrophobic clusters. In a more sophisticated way, they could integrate several outputs to produce a consensus by some predefined criterion and cover more aspects of the disorder, ([39]) ([30]; [31]).

Certain disadvantages in epitope-based vaccines are due to low immunogenicity and difficulties related to the fine identification of protective epitopes and/or properly folded antigen structural motifs to be included in a vaccinal preparation. The latter is fundamental to properly activate an effective immune response and strenghts the significance of natural processing and presentation of Ags. The aim of this research is to determine whether disorder- and hydropathy- prediction methods could help in understanding selective T-cell epitope processing and presentation. Since some disorder predictors can capture different types and functional aspects of the disorder, the primary task was to demonstrate if the epitope prevalent location in structural order is valid for various disorder predictors. The relationship between epitope prediction (epitope frequency and binding affinity), hydropathy and distribution of epitopes within predicted ordered and disordered protein regions was analyzed. Both HLA class-I and HLA class-II binding epitopes were found to be more frequent and prevalently hydrophobic in ordered than in the disordered protein regions ([32]). This paper provides the continuation of the previous research. We focus on the analysing epitopes in the consensus of 9 disorder predictors, since the consensus of order or disorder scores of different predictors covers more aspects of the disorder and provides, with high probability,stricter criteria of disorder. It also eliminates the influence of different disorder-prediction algorithms on epitope characteristics.

2 Methods

2.1 Data

The research was performed on the same material as in the research described in [32]. The database containing 619 proteins. Most proteins (465) were downloaded from the DisProt database [33], release 5.0 , which contains proteins with experimentally determined regions of disordered structure, originating from various organisms. Other proteins are selected based on two criteria:

- to balance the number of eukaryotic/prokaryotic as well as ordered/disordered proteins. As proteins from DisProt are over 70% eukaryotic and disordered, we took, as compensation, additional 115 prokaryotic proteins (downloaded from the National Center for Biotechnology Information [34]), and 15 proteins with experimentally proved order structure (downloaded from PDB database [35]).
- to form a group of (human) tumor-associated antigens (TAA). A group of 19 most investigated CTAs and five additional TAA was chosen.

All sequences have been assigned to one of taxonomic categories (223 bacteria, 376 eukarya, and 20 eukaryotic viruses).

2.2 Epitope and Disored prediction

For the epitope prediction we have used the well-characterized NetMHCpan-2.0 and NetMHCIIpan-1.0 programs from CBS [36]. NetMHCpan program generates quantitative predictions of the affinity of any peptide-MHC class-I interaction. It covers HLA-A, B, C (Cw), D and E and alleles from several non-human species. NetMHCIIpan generates predictions for MHC class-II HLA-DRB alleles. We have chosen nonamer peptide sequence for our analysis since most HLA molecules have a strong preference for binding 9mers. The number of available human alleles in NetMHCpan-2.0 and NetMHCIIpan-1.0 programs was 1469 and 517, respectively. Since some alleles have identical pseudo-sequences, we have chosen HLA allele, whose name was the first in alphabetical order, as the representative of each of these allelic groups. After exclusion, 1006 different HLA class-I and 326 different HLA class-II alleles exist with unique pseudo-sequences. Among them, 45 HLA-I alleles do not produce any (either SB or WB) epitope. The overall numbers of predicted epitopes used as an input in this research were

2037890 (HLA-I) and 1065338 (HLA-II). Together with epitopes binding strength in the research we have used supertypes related to corresponding alleles. Both HLA alleles groups were classified according two different classifications: HLA-I allele are classified according to [27] and [37], while HLA-II allele are classified according to [38] and [37].

In order to eliminate the impact of an individual disorder prediction algorithm, disordered and ordered protein regions have been predicted in two steps:

- In the first step we have used 7 different disorder predictors with the total of 9 variants (Table 3). Predictors have been chosen according to the following criteria: (1) predictor is freely available and can be downloaded and executed in local environment, (2) predicting disordered regions is not a long-running task, and (3) predictors are based on different prediction methods. Several of these disorder predictors were found to be among the top methods, according to CASP experiments (Critical Assessment of protein Structure Prediction, [39], [40]), or comparable to these methods ([41]).

Predictor	Source location
VSL2b	http://www.ist.temple.edu/disprot/predictorVSL2.php
IsUnstruct V2.02	http://bioinfo.protres.ru/IsUnstruct/
IuPred 1.0 - long disorder (IuPred -Long) - short disorder (IuPred-Short)	http://iupred.enzim.hu/
RONN 3.1	http://www.strubi.ox.ac.uk/RONN
Disembl - Hot-loops - Remark465	http://dis.embl.de/
OnDCRF 1.0	http://babel.ucmp.umu.se/ond-crf/
DISOPRED 2.4.3	http://bioinf.cs.ucl.ac.uk/psipred/?disopred=1

Table 3: Disorder predictors used in the research

- In the second step, for every individual protein a consensus of predicted regions have been determined. Consensus is considered as intersection of positions in protein of the predicted regions with the same type (ordered or disordered) for every used disordered predictor.

By associating results of epitope and disorder predictions, it can be seen that epitopes can be classified into three groups. The first one (named O epitopes) includes epitopes completely belonging to the consensus of predicted ordered regions; the second (named D epitopes) includes epitopes completely belonging to consensus of predicted disordered regions; and the third one (named N epitopes) includes epitopes that do not belong to any consensus of predicted ordered/disordered regions.

Analysis of the distribution and number of epitopes has been performed both in terms of absolute values and values normalized to 100AA, related to the length of the corresponding region type. Available interval in the region in which epitope can appear is equal to the sum of the regions length for D and O epitopes, for N epitopes it is equal (total_protein_length - length_for_O_epitopes - length_for_D_epitopes), while for B epitopes, it is equal to:

- 16 AA, if the region is not on protein termini and is longer than 16 AA (16 AA, because there is a possible interval of 8 AA at each end of the region);
- $min(8, length(region))$ AA, if the region is located on the protein terminal. If the region is longer than 8 AA, then only 8 AA on non-terminal part can host the N epitope. If the region is shorter than 8 AA, it completely fits into the available interval;
- otherwise, the length of the region.

Analysis in terms of HLA supertypes have been perfomed related to the normalized (100AA) values, to the number of alleles in the supertype, or both.

2.3 Hydropathy prediction

Two of the most commonly used hydrophobicity scales: Kyte-Doolittle (KD, [42]) and Hopp-Woods (HW, [43]) scales, were applied for the prediction of the hydropathy plots of analyzed proteins and appertaining epitopes. The HW scale is a hydrophilic index and defines 11/20 AA acids as hydrophobic, while the KD scale is limited to 7/20 most hydrophobic AA, meaning that the Kyte-Doolittle scale is stricter than the Hopp-Woods scale, regarding the individual AA hydrophobicity definition. Hydropathy of individual proteins has been calculated using a sliding window and by summing up scores from standard AA hydrophobicity scales. The sliding window size can be varied according to the expected size of the structural

motif under investigation. We have opted for the window of 9 AA because of the chosen peptide length of 9 AA for the epitope prediction. A peptide hydropathy can be determined in two ways: as the average hydropathy value (AvgH) of contained AA, and by counting the number of hydrophobic/hydrophilic AA in the epitope (named majority of AA, MAA). In the AvgH method, we have assumed that peptide would be considered as hydrophobic, if an average hydropathy was ≥ 0 for the KD scale, or ≤ 0 for the HW scale. In the MAA method, peptide would be considered as hydrophobic, if a majority of AA were hydrophobic (for both scales). Peptides not considered hydrophobic were considered hydrophilic.

3 Results

3.1 Proteins, disorder and hydropathy

Table 2. captures data on proteins belonging to the main taxonomic categories (bacteria, eukarya and eukaryotic viruses). The usage of hydrophobic AA has varied from 32.92% in eukarya, to 40.41% in bacteria (Kyte-Doolittle scale), and from 43.34% in eukaryotic viruses to 51.66% in bacteria (Hopp-Woods scale) and corresponded to what was previously found on lower median usage of hydrophobic AA in eukarya (36.4-38.4%) as opposed to bacteria (38.7-43.5%) ([44]). The percentage of disordered regions

Taxon	Number of proteins	Protein length (AA)			% hydrophobic AA content	
		Minimal	Average	Maximal	KD scale	HW scale
bacteria	223	21	246,21	1091	40,41	51,66
eukarya	376	32	358,59	1685	32,92	44,28
viruses_euk	20	71	335,50	641	33,16	43,34
complete mat.	619	21	317,36	1685	35,02	46,31

Table 4: Protein characteristics, grouped according to taxonomic categories

of various lengths in proteins from the complete material varied from 21% to 48%, depending on the disorder predictor, regardless of the length of disordered regions. The percentage of long consecutive disordered regions ($\geq$30 AA) in analyzed proteins roughly corresponds to the previously published results concerning the content of disordered regions in bacteria and eukarya superkingdoms. Dunker and co-workers ([45]) have found that the percentage of long disordered regions ($\geq$30 AA, $\geq$40 AA or $\geq$50 AA), predicted by the PONDR-VLXT predictor, was significantly higher in eukarya than in bacteria. Similar results were obtained using VL2 predictors for re-

gions >40 AA: 2%-38% in bacteria and 31%-53% in eukarya ([46]). We have found that the disorder content of proteins from eukaryotic viruses is similar to the disorder content of the corresponding hosts (eukarya), although the result has to be taken with caution, due to the small number of analyzed viral proteins.

Distribution of regions and hydrophobicity over proteins have similar trends for all disorder predictors. As expected, number of hydrophobic AA was higher if counted according to HW scale. For the analysis of the distribution of predicted regions in our protein dataset, the hypothetical protein and region were constructed, by normalization of the length of all proteins and regions to 100. Normalization is performed by elongating the proteins to the longest protein length (1685) and then resizing it to 100. Similarly, predicted regions were extended to the longest region length for each predictor (for example, for VSL2b predictor lengths are 799 for D regions and 528 for O regions) and then resizing it to 100. Figure 1 show distribution of number of regions with various lengths and AA hydrophobicity over protein. Predictions were made by VSL2b predictor and KD hydrophobicity scale. It is evident that the distribution of hydrophobic/hydrophilic AA do not depends on their position in the protein, with exception of the protein N-terminal. This distribution is caused by the prevalence of AA Methionine in the N-terminal. Although Methionine is the first amino acid of newly synthe- sized proteins in majority of organisms, it is usually removed from mature pro- teins to leave a nonbulky N-terminal residue. The figure also shows that protein N- and C-terminals are enriched in (relatively short) disordered regions. As distance from the protein ends increases, the number of disordered regions decreases. Approximately at 20% (80%) of protein length the number of ordered regions exceeds the number of disordered regions. As protein ends are characterized by short regions, consensus regions are usually found in the middle part of the protein, particularly those longer than 9 AA.

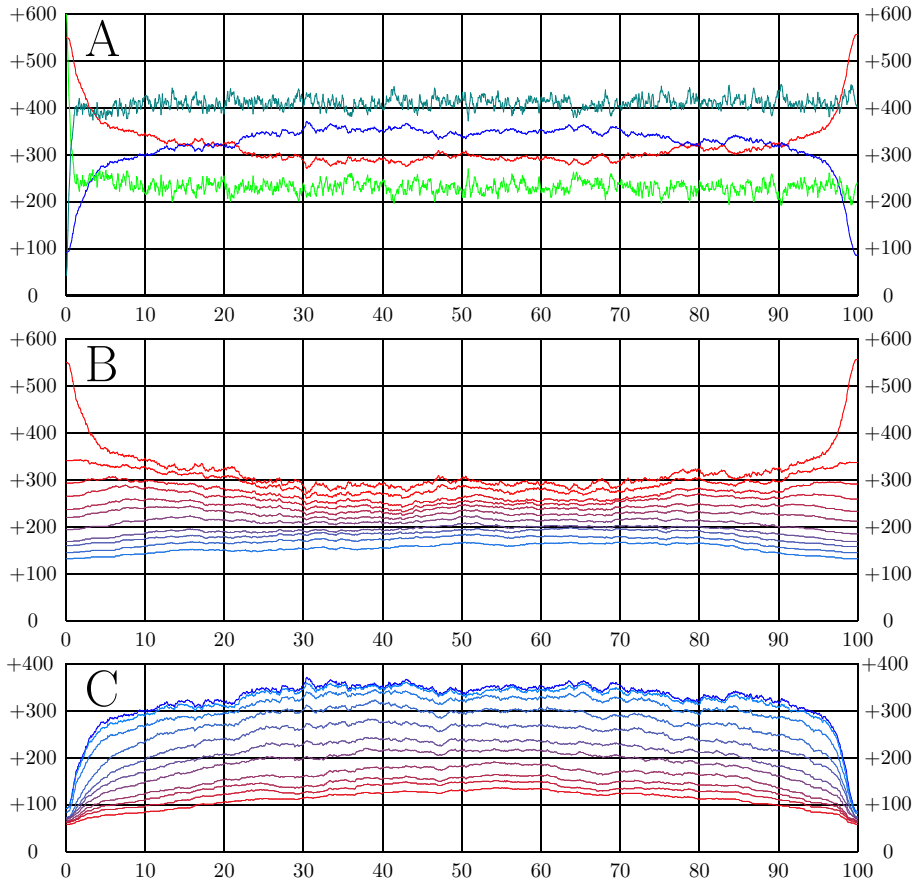


Figure 1: Distributions of regions and hydrophobicity over proteins. Part A: Number of D regions, O regions, hydrophobic and hydrophilic AA

Part B: Distributions of disorder regions over proteins. length ≥1, length ≥10, length ≥20, length ≥30, length ≥40, length ≥50, length ≥60, length ≥70, length ≥80, length ≥90 and length ≥100

Part C: Distributions of order regions over proteins. length≥1, length≥10, length≥20, length≥30, length≥40, length≥50, length≥60, length≥70, length≥80, length≥90 and length≥100

Taxon	% of disordered regions	% of disordered regions with length≥30	% of disordered regions with length≥30
bacteria	5,26	1,06	0,81
eukarya	11,95	3,61	2,47
viruses_euk	14,11	8,33	8,33
complete mat.	10,15	3,06	2,21

Table 5: Number and percent of disordered regions in consensus

Consensus of predicted regions

Approximately 25% of all predicted regions, defined by some individual predictor fall into a set of consensus regions related to all predictors. Percent of consensus of disordered regions rapidly decreases with increasing of region length (table 5), which is the consequence of similar trend in both DisEMBL predictors. Among them, 35% of consensus of disordered regions and 60% of the consensus ordered regions have the length ≥10 which provides sufficient material for the analysis of epitopes (with length of 9AA) in such regions. Average region lengths of consensus regions were 9.7 (for disordered) and 20.6 (for ordered) regions.

There are significant differences in AA contents of all predicted regions (related to individual disorder predictor) and the consensus of predicted regions. Amino acids C, F, I, L, V, W and Y are underrepresented in the consensus of disordered regions and overrepresented in the consensus ordered regions, while amino acids E, K, R and S are overrepresented in the consensus disordered regions and underrepresented in the consensus of ordered regions, compared to their content in all regions (figure 2). As a consequence, if region predicted to be disordered (or ordered) by a individual disorder predictor have AAs content similar to the previously described relations, it will probably be predicted as disordered (or ordered) if prediction is performed by some other disorder predictor.

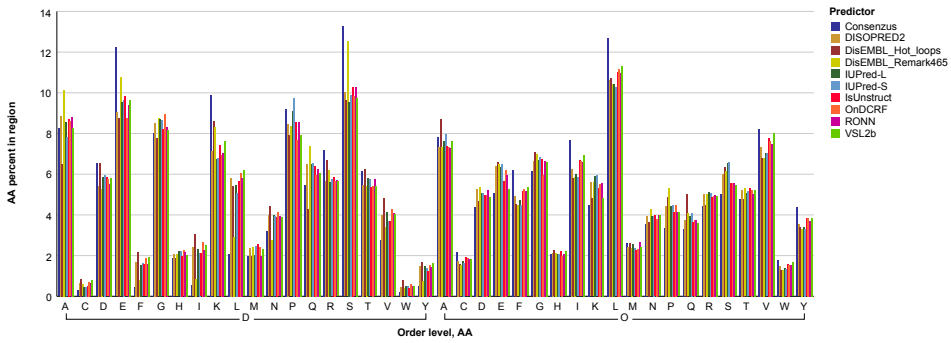


Figure 2: Comparison of AA percent in consensus/all regions predicted by used set of disorder predictors

Epitopes in consensus regions

The number of epitopes/100 AA and its percent in transitional regions, encompassing consensuses of ordered regions was calculated related to length of corresponding regions type. Length was calculated as descirbe on page 131. Epitopes in consensus regions have different amino acids contents compared to consensus regions they belongs to. Table 6 contains the percentage of the deviation in the number of AAs in epitopes related to the region type.

The most distinctive characteristics of epitopes in all regions are that AAs cysteine, aspartic acid, glutamine and glutamic acid are underrepresented in epitopes, compared to the corresponding regions, while AAs leucine and serine are overrepresented in epitopes, compared to its content in regions. AA valine is (highly) overrepresented in disordered regions, etc. Table Table 6 can be used as the basis for giving some directions if some areas in regions can be located as a proper source of epitopes. Rules can be defined based on the percentage of AA in some part and/or in corresponding consensus region. For example:

- if a part of D region has higher percentage of AA leucine or serine compared to the percentage of this AA in the region, then, a higher probability exists that some (HLA-I or HLA-II) epitopes will be located inside such region;

- if the part of D region has a higher percentage of AA threonine, compared to the percentage of this AA in the region, then a higher

AA_PERCENT		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
Consensus reg.	D	8.25	0.29	6.53	12.23	0.43	8.04	1.88	0.53	9.88	2.09	1.96	3.19	9.2	5.44	7.19	13.25	6.17	2.75	0.19	0.51
	O	7.82	2.16	4.38	5.06	6.2	6.14	2.07	7.66	4.48	12.69	2.62	3.56	3.34	3.31	4.42	5	4.75	8.21	1.77	4.36
HLA-I	D	11.69	0.14	1.45	5.73	0.99	4.47	1.99	0.51	11.19	3.01	1.32	1.92	10.22	3.89	10.79	17.12	8.31	3.93	0.26	0.98
	O	8.69	1.29	2.17	3.33	9.52	4.71	1.69	8.42	3.1	14.78	3.86	2.66	3.53	2.66	3.42	5.85	4.63	7.85	2.13	5.61
HLA-II	D	7.24	0.14	1.3	2.72	0.9	4.74	1.54	0.95	8.65	6.5	2.27	3.54	7.15	3.2	10.12	25.09	6.65	5.74	0.19	1.25
	O	7.36	0.84	1.73	1.84	6.22	5.06	1.46	9.98	3.53	19.06	3.7	3.68	2.54	2.02	3.96	7.69	4.6	9.68	1.22	3.76

Table 6: Content of amino acids in epitopes in disordered (D) and ordered (O)regions, compared to their content in the corresponding consensus regions. Values for overrepresented AA in epitopes, compared to the content of the same AA in the corresponding region are shown in filled dark green box (20% or more higher than in regions) and in filled light green box (between 5% and 20% higher than in regions). Values for underrepresented AA in epitopes compared to the content of the same AA in O or N region are shown in filled orange box (between 5% and 20% lower than in regions) and filled dark red box (20% or more percent lower than in regions). Transparent (white) boxes denotes values with no significant distinction (less than 5%) from the percentage of AA in the consensus region. Filled gray boxes - percentage of an AA in O and D consensus regions (presented for comparison). The one letter AA code is used.

probability exists that some HLA-I epitopes will be located inside such a region;

- if part of the ordered region has a lower percentage of AA lysine, this area is a potential source of epitopes;
- if a part of a region has a higher percentage of AA proline, than the region itself, this area includes HLA class-I epitopes; but if it has a lower percentage of AA proline, this area is a potential source of HLA-II epitopes;
- etc.

Note: previous rules are affirmative, not negative. For example, it is not prohibited that HLA-I epitope can be found in the area with a low percentage of AA proline.

Hydrophobicity of epitopes in consensus regions

The data, obtained on all proteins in the database, revealed that both HLA-1 and HLA-II epitopes concentrate in the consensus of ordered regions, as

was previously shown for 19 tumor associated Ags from the cancer-testis Ag group ([32]). However, grouping of epitopes in transitional regions, on the borderlines of the consensuses of ordered regions, indicates the potential significance of the borderline regions for epitope structural availability ([11]). Figure 3 represents frequency, affinity and hydropathy of epitopes in D, O, B and N region. For all hydropathy scales and in all types of region SB epitopes are prevalently hydrophobic; on the other side, although in the majority of region/scale combination WB epitopes are also prevalently hydrophobic, exists some regions where WB epitopes are prevalently hydrophilic. This also can be used for determining epitope characteristics and direction where to find epitope of certain class.

If epitope is considered as a part of supertype according to corresponding allele, it can be seen from figure 3 that majority of SB epitope bound per single allele is concentrated in ordered consensus regions. The borderline (B) regions (which partially cover some consensus region) have higher percent of epitopes than N-type of regions which do not include any consensus region. Simliar relations hold for WB epitpoes, although exists supertypes (for example, B7/Multipred1).

4 Experimental evidences on T-cell epitope location in ordered and disordered consensus regions of cancer -testis Ag MAGE A3

Cancer-testis Ags (CTA), is a group of antigens expressed in various tumors, but silenced in normal cells ([47]). We have studied distriution of HLA-I and HLA-II epitopes, predicted y Net-MHC programs and experimentally validated ([52]) in consensus of O/D regions. Majority of epitopes, for the 19 CTAs was found in areas encompassing the consensus of O regions, predicted by all used disorder predictors, or in their close surrounding.

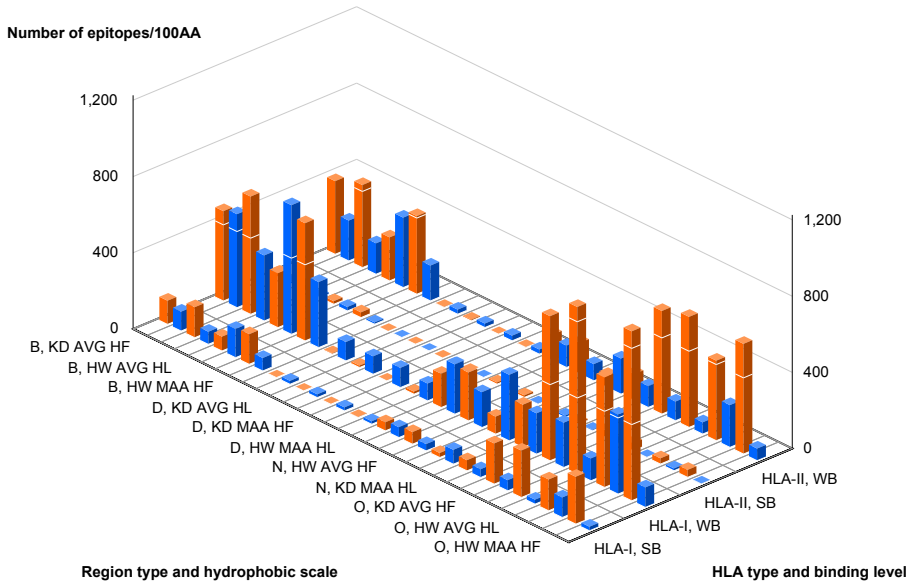


Figure 3: Frequency, affinity and hydropathy of epitopes in D, O, B and N regions appearing in the complete protein material, determined for all analyzed HLA class I and II alleles. Hydropathy is presented over KD and HW scales (AVG and MAA methods). Columns that represent hydrophobic epitopes are colored in orange; columns that represent hydrophilic epitopes are colored in blue. X-axis represents the region type and hydrophobic scale. On the x-axis in figure, every 4th label is marked. Labels are ordered according scheme B-D-N-O for region type, and KD AVG HF, KD AVG HL, HW AVG HF, HW AVG HL, KD MAA HF, KD MAA HL, HW MAA HF, HW MAA HL for hydrophobicity.

A CTA MAGE-A3 (UniProt Acc No: P43357), expressed in a broad spectrum of malignancies, was intensively studied for tumor immunotherapy ([48]; A). We have observed clustering of predicted HLA-I and HLA-II epitopes in the C-terminal region of MAGE-A3, which encompass the consensus of O protein regions, defined by all used disorder predictors, figure 5A. It is evident that the gaps in the predicted consensus of O regions coincide with the minimums of predicted epitopes, particularly those of HLA-II type. Atanackovic and colleagues ([49]) observed that HLA-I and HLA-II epitopes (both previously described and those obtained in their study) were clustered in the C terminal region of the protein. On the contrary, the re-

gion, predicted to be disordered by all used disorder predictors (consensus of D regions) is almost completely depleted of experimentally verified epitopes, figure 5. Besides, the response to the naturally processed HLA-II epitopes from MAGE-A3 was found to be promiscuous for several DRB1, DQB1 or DPB1 alleles and located in ordered regions (which are also prevalently hydrophobic) ([32]). Summarized results on experimentally verified MAGE-A3 epitopes revealed that they are located in the consensus of ordered protein regions and are prevalently hydrophobic, for both scales and methods, except for KD MAA method. Experimentally verified epitopes are evidently restricted to the area that is, more narrow than the area of the predicted epitopes, which may have implications on the rational vaccine design.

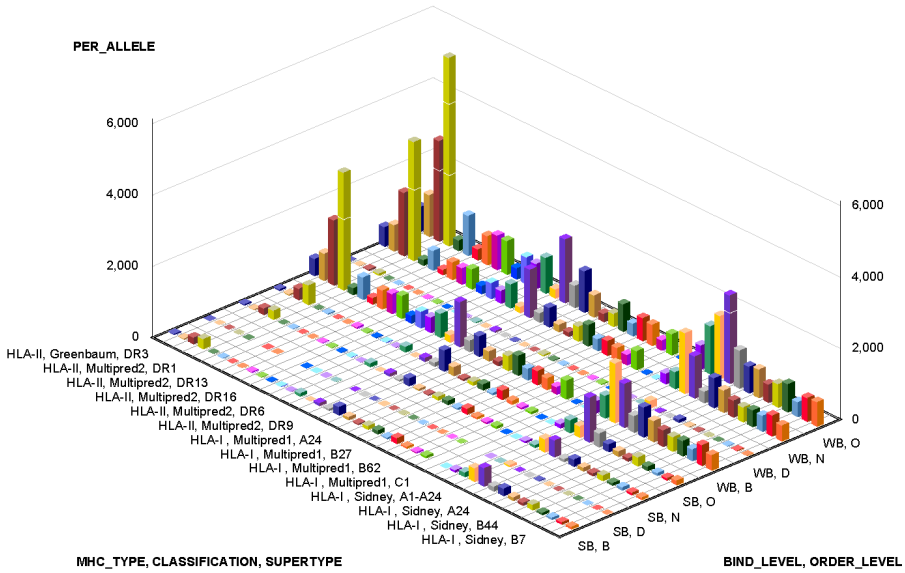


Figure 4: Numner of epitopes per allele/supertype in concesus region. Multipred1 and Multipred2 are classification od HLA-I and HLA-II alleles according to [37]; Sidney is classification of HLA-I alleles according to [27], and Greenbaum is classification of HLA-II alleles according to [38]

T-cell immune response to CTA, which escapes to the central and peripheral tolerance, is supposed to correspond to those against the non-self Ags. An immune response against the non-self Ags, is driven by the T-cell receptor affinity for the peptide-MHC complex and is selected to attack

the strongest antigenic part(s) of pathogens ([50]). Under selective pressure, the strongest antigenic parts of pathogenic (or non-self) molecules, will become those that are functionally important and will be recognized by different MHC alleles. Intrinsic characteristics, such as the prevalence of hydrophilic or hydrophobic AA residues and secondary structural elements, could mainly affect peptide binding to the MHC molecules. However the superposition of prevalently hydrophobic peptides, predicted to bind with high affinity to MHC molecules with the structurally favorable position, will positively influence peptide processing and presentation, at least in proteins with limited 3D structure, that can easily be unfolded, or those that are not heavily glycozylated, ([51]).

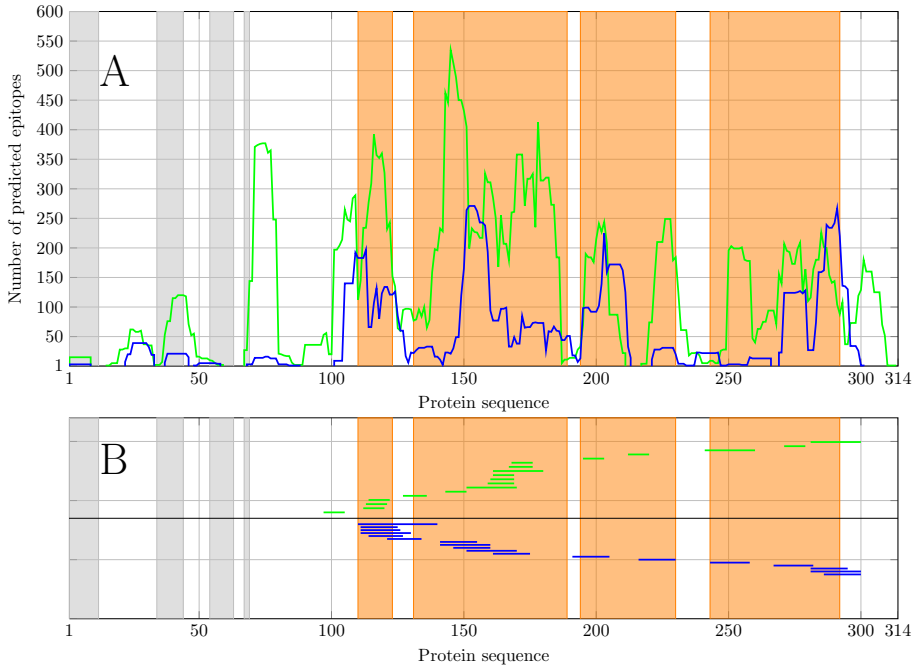


Figure 5: Distribution of epitopes in human MAGE-A3 protein (UniProtKB/Swiss-Prot Acc. No. P43357). A) Epitopes predicted by NetMHCpan (HLA class-I) (green) and NetMHCIIpan methods (HLA class-II) (blue). HLA binding predictions are subjected to the total number of 1006 HLA class-I and 326 HLA class-II alleles, with unique pseudosequences. B) Experimentally validated HLA class-I (green) and HLA class-II (blue) binding epitopes ([52]). Consensus is coloured orange (ordered regions) and grey (disordered regions).

5 Conclusion

T-cell epitope prediction, based on antigen-derived peptide affinity to MHC molecules, does not cover additional aspects of immunogenicity, such as the influence of epitope position within the Ag 3D structure on its availability to protease cleavage and presentation to T-cells, which might influence epitope immunodominance. The motivation for analyzing the relationship between an epitope hydropathy and affinity and the distrbution of epitopes within predicted structural regions was to find out if prospective epitopes have structural restrictions. Disordered protein regions are prevalently hydrophilic and extremely sensitive to proteolysis in vitro, and are expected to be under-represented as T-cell epitopes. We particularly considered the epitopes predicted to bind with high affinity to HLA molecules, which are of interest in vaccine studies, regarding their distribution in consensuses of order or disordered protein regions, predicted by different disorder predictors. The consesus of structural regions covers more aspects of disorder eliminates the influence of different disorder-prediction algorithms on epitope characteristics. The frequency of HLA class-I or -II epitopes, predicted by the pan-specific T-cell epitope predictors NetMHCpan and NetMHCIIpan in the consensus of of 9 publicly available disorder predictors was found to be almost 10 times higher in consensus of ordered than in consensus disordered protein regions. Both HLA class-I and HLA class-II binding epitopes are prevalently hydrophilic in disordered and prevalently hydrophobic in ordered protein regions, whereas epitopes recognized by HLA class-II alleles were more hydrophobic than those recognized by HLA class-I. As regards both classes of HLA molecules, high-affinity binding epitopes display more hydrophobicity than low affinity-binding epitopes (in both ordered and disordered regions), meaning that prospective vaccine candidates are prevalently hydrophobic. Epitopes belonging to disordered protein regions were not predicted to have poor affinity to HLA class-II molecules, as expected from disorder intrinsic proteolytic instability. The relation of epitope hydrophobicity and order/disorder location was also valid if alleles were grouped according to the HLA class-I and HLA class-II supertypes, These data suggests that reverse vaccinology, oriented towards high-affinity epitopes, is also oriented towards prevalently hydrophobic epitopes encompassing the consensus of ordered regions. The analysis of predicted and experimentally evaluated epitopes of cancer-testis antigen MAGE-A3 has confirmed that the majority of T-cell epitopes, particularly those that are or naturally processed, was located in ordered and disorder/order boundary protein regions overlapping hydrophobic regions.

Acknowledgements

This work is supported by the Ministry of Education, Science and Technological Development, Republic of Serbia, Projects No. 174021, 174002, and III44006.

References

[1] V. Brusic, V. B. Bajic and N. Petrovsky,*Computational methods for prediction of T-cell epitopes–a framework for modelling, testing, and applications*, Methods, *34(4), 436-443*, (2004)

[2] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz and V. Brusic, *Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research*, *BMC Bioinformatics, 9 Suppl 12, S22*, (2008)

[3] H. H. Lin, S. Ray, S. Tongchusak, E. L. Reinherz and V. Brusic, V. *Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research*, *BMC Immunol, 9, 8*, (2008)

[4] X. Yang and X. Yu, *An introduction to epitope prediction methods and software*, *Rev Med Virol, 19(2), 77-96*, (2009)

[5] O. Lund et al., *Definition of supertypes for HLA molecules using clustering of specificity matrices*, Immunogenetics, *55(12), 797-810*, (2004)

[6] A.J. Bordner, *Towards universal structure-based prediction of class II MHC epitopes for diverse allotypes*, *PLoS One, 5(12), e14383*, (2010)

[7] A. Patronov and I. Doytchinova, *T-cell epitope vaccine design by immunoinformatics*, *Open Biol, 3(1), 120139*, (2013)

[8] M. Nielsen et al., *NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence*, *PLoS One, 2(8), e796*, (2007)

[9] M. Nielsen et al., *Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan*, *PLoS Comput Biol, 4(7), e1000107*, (2008)

[10] A. Sette and J. Fikes, *Epitope-based vaccines: an update on epitope identification, vaccine design and delivery*, *Curr Opin Immunol, 15(4), 461-470*, (2003)

[11] S. J. Landry, *Local protein instability predictive of helper T-cell epitopes.* Immunol *Today, 18(11), 527-532*, (1997)

[12] G. Dai, S. Carmicle, N. K. Steede and S. J. Landry, *Structural basis for helper T-cell and antibody epitope immunodominance in bacteriophage T4 Hsp10. Role of disordered loops*, *J Biol Chem, 277(1), 161-168*, (2002)

[13] S. Carmicle, N. K. Steede and S. J. Landry, *Antigen three-dimensional structure guides the processing and presentation of helper T-cell epitopes*, *Mol Immunol, 44(6), 1159-1168*, (2007)

[14] D. Mirano-Bascos, M. Tary-Lehmann and S. J. Landry, *Antigen structure influences helper T-cell epitope dominance in the human immune response to HIV envelope glycoprotein gp120*, *Eur J Immunol, 38(5), 1231-1237*, (2008)

[15] S. J. Landry, *Helper T-cell epitope immunodominance associated with structurally stable segments of hen egg lysozyme and HIV gp120*, *J Theor Biol, 203(3), 189-201*, (2000)

[16] S. J. Melton and S. J. Landry, *Three dimensional structure directs T-cell epitope dominance associated with allergy*, *Clin Mol Allergy, 6, 9*, (2008)

[17] K. A. Chianese-Bullock et al., *Antigen processing of two H2-IEd-restricted epitopes is differentially influenced by the structural changes in a viral glycoprotein*, *J Immunol, 161(4), 1599-1607*, (1998)

[18] J. A. Musson et al., *Differential processing of CD4 T-cell epitopes from the protective antigen of Bacillus anthracis*, *J Biol Chem, 278(52), 52425-52431*, (2003)

[19] J. M. Weaver et al., *Immunodominance of CD4 T cells to foreign antigens is peptide intrinsic and independent of molecular context: implications for vaccine design*, *J Immunol, 181(5), 3039-3048*, (2008)

[20] J. M. Weaver and A. J. Sant, *Understanding the focused CD4 T cell response to antigen and pathogenic organisms*, *Immunol Res, 45(2-3), 123-143*, (2009)

[21] A. J. Sant et al. *The relationship between immunodominance, DM editing, and the kinetic stability of MHC class II:peptide complexes*, Immunol Rev, 207, 261-278, (2005)

[22] Uversky, V. N., Gillespie, J. R., Fink, A. L. *Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins, 41(3), 415-427*,(2000) .

[23] Uversky, V. N. *The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. J Biomed Biotechnol, 568068. doi: 10.1155/2010/568068* (2010).

[24] Uversky, V. N.,Dunker, A. K. *Understanding protein non-folding. Biochim Biophys Acta, 1804(6), 1231-1264. doi: 10.1016/j.bbapap.2010.01.017* ,(2010).

[25] Carl, P. L., Temple, B. R., Cohen, P. L. *Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity. Arthritis Res Ther, 7(6), R1360-1374. doi: 10.1186/ar1832*, (2005).

[26] Suskiewicz, M. J., Sussman, J. L., Silman, I., Shaul, Y. *Context-dependent resistance to proteolysis of intrinsically disordered proteins.Protein Sci, 20(8), 1285-1297. doi: 10.1002/pro.657*,(2011).

[27] Sidney, J., Peters, B., Frahm, N., Brander, C., Sette, A. *HLA class I supertypes: a revised and updated classification. BMC Immunol, 9, 1. doi: 10.1186/1471-2172-9-1*,(2008).

[28] Halling-Brown, M., Shaban, R., Frampton, D., Sansom, C. E., Davies, M., Flower, D., Duffield, M., Titball, R. W., Brusic, V., Moss, D. S. *Proteins accessible to immune surveillance show significant T-cell epitope depletion: Implications for vaccine design.Mol Immunol, 46(13), 2699-2705. doi: 10.1016/j.molimm.2009.05.027*,(2009).

[29] Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., Dunker, A. K. *Intrinsic disorder and functional proteomics. Biophys J, 92(5), 1439-1456. doi: 10.1529/biophysj.106.094045*,(2007).

[30] Dosztanyi, Z., Meszaros, B., Simon, I. *Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins.Brief Bioinform, 11(2), 225-243. doi: 10.1093/bib/bbp061*,(2010).

[31] Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., Rost, B. *Improved disorder prediction by combination of orthogonal approaches. PLoS One, 4(2), e4433. doi: 10.1371/journal.pone.0004433*,(2009).

[32] N. S. Mitić, M. D. Pavlović, D. R. Jandrlić, *Epitope distribution in ordered and disordered protein regions part A. T-cell epitope frequency, affinity and hydropathy*, accepted for publication in *Journal of Immunological Methods*

[33] Database of Protein Disorder: http://www.disprot.org

[34] National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov

[35] RSCB Protein Data Bank: http://www.pdb.org

[36] The Center for Biological Sequence Analysis at the Technical University of Denmark: http://www.cbs.dtu.dk/services/

[37] G. L. Zhang et al., *MULTIPRED2: a computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles*, J Immunol Methods, *374(1-2), 53-61*, (2011)

[38] J. Greenbaum, J. Sidney, J. Chung, C. Brander, B. Peters and A. Sette, *Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes*, Immunogenetics, *63(6), 325-335*, (2011)

[39] P. Tompa, *Structure and Function of Intrinsically Disordered Proteins*, Chapman and Hall/CRC, (2009)

[40] O. Noivirt-Brik, J. Prilusky and J. L. Sussman, *Assessment of disorder predictions in CASP8*, *Proteins, 77 Suppl 9, 210-216*, (2009)

[41] M. Y. Lobanov and O. V. Galzitskaya, *The Ising model for prediction of disordered residues from protein sequence alone*, Phys Biol, 8(3) 035004, (2011)

[42] J. Kyte and R. F. Doolittle, *A simple method for displaying the hydropathic character of a protein*, J Mol Biol, 157(1), 105-132, (1982)

[43] T. P. Hopp and K. R. Woods, *Prediction of protein antigenic determinants from amino acid sequences*, Proc Natl Acad Sci USA, 78(6), 3824-3828, (1981)

[44] S. Karlin, L. Brocchieri, J. Trent, B.E. Blaisdell and J. Mrazek, *Heterogeneity of genome and proteome content in bacteria, archaea, and eukaryotes*, Theor Popul Biol, 61(4), 367-390, (2002)

[45] A. K. Dunker, Z. Obradovic, P. Romero, E. C. Garner and C. J. Brown, *Intrinsic protein disorder in complete genomes*, Genome Inform Ser Workshop Genome Inform, 11, 161-171, (2000)

[46] Vucetic, S., Brown, C. J., Dunker, A. K., Obradovic, Z. *Flavors of protein disorder.*Proteins, 52(4), 573-584. doi: 10.1002/prot.10437,(2003).

[47] Scanlan, M. J., Gure, A. O., Jungbluth, A. A., Old, L. J., Chen, Y. T. *Cancer/testis antigens: an expanding family of targets for cancer immunotherapy.*Immunol Rev, 188, 22-32. ,(2002).

[48] Straetemans, T., van Brakel, M., van Steenbergen, S., Broertjes, M., Drexhage, J., Hegmans, J., Lambrecht, B. N., Lamers, C., Bruggen, P. G., Coulie, P. G., Debets, R. *TCR gene transfer: MAGE-C2/HLA-A2 and MAGE-A3/HLA-DP4 epitopes as melanoma-specific immune targets.*Clin Dev Immunol, 2012, 586314. doi: 10.1155/2012/586314 ,(2012).

[49] Atanackovic, D., Altorki, N. K., Cao, Y., Ritter, E., Ferrara, C. A., Ritter, G.,Hoffman, E.W., Bokemeyer, C., Old, L. J., Gnjatic, S. *Booster vaccination of cancer patients with MAGE-A3 protein reveals long-term immunological memory or tolerance depending on priming.*Proc Natl Acad Sci U S A, 105(5), 1650-1655. doi: 10.1073/pnas.0707140104,(2008).

[50] Dai, Y. D., Carayanniotis, G., Sercarz, E. *Antigen processing by autoreactive B cells promotes determinant spreading.* Cell Mol Immunol, 2(3), 169-175.,(2005).

[51] Brown, S. A., Stambas, J., Zhan, X., Slobod, K. S., Coleclough, C., Zirkel, A., Surman, S., White, S. W., Doherty, P. C., Hurwitz, J. L. *Clustering of Th cell epitopes on exposed regions of HIV envelope despite defects in antibody activity.*J Immunol, 171(8), 4140-4148.,(2003).

[52] Cancer immunity database (http://www.cancerimmunity.org/)

Local Protein Structure Prediction by Bayesian Probabilistic Approach Principle

Mihajlo Mudrinić[a]
Institute of Nuclear Sciences "Vinča", P.O.Box 522, 11001 Belgrade, Serbia

Abstract

The task of understanding and predicting how to translate the information coded in the amino acid sequence of proteins into knowledge of how such protein would fold, is one of the most important problems in biochemistry. Consequently, we want to understand how the primary structure (the sequence of residues) gives rise to tertiary structure (the folded state). The work is focused on intermediate structure between the two, the secondary structure (patterns or motifs like helices). First step known as data fusion in methodology for visualization will be presented. Data fusion entails integration of paired analysis (based on the relation protein block-amino acid propensity) into knowledge based on estimation theory including statistical pattern recognition and multivariate analysis.

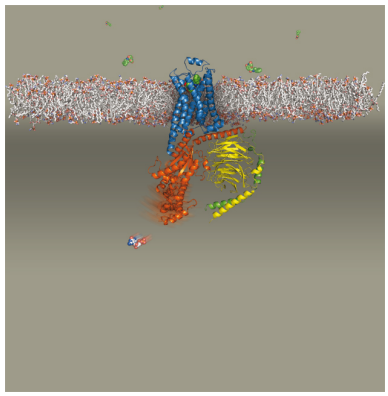
[a] e-mail address: m.mudrinic@vinca.rs

1 Introduction

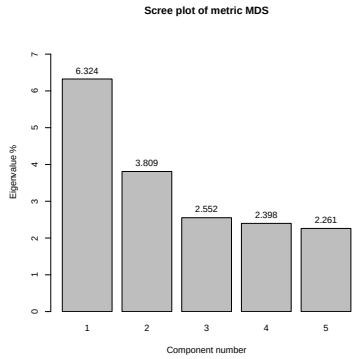
With an increasing volume and types of biological data, data fusion as integration of various types of protein-reated data is becoming an important part in research leading to identfication of homology in protein structure classification. It has been demonstrated [1] that incorporating knowledge derived from amino acid sequences and known protein-protein interactions improves classification performance compared to any single type of data. Proteins are very important elements of life, for example even the sensation of exogenous stimuli, such as light, odors, and taste, is mediated via the largest superfamily of proteins G-Protein Coupled Receptors (GPCRs). GPCRs are vital protein bundles with their key role in cellular signaling and regulation of various basic physiological processes. These properties of GPCRs make them an excellent potential therapeutic target class for drug design. Visualization of G protein-coupled receptors which are membrane proteins with a beta sheet and several alpha helices is given on figure 1(a). The information useful for analyzing genetic relatedness and the sequence-function relationships of protein families is the degree of similarity between sequences of amino acids. The degree of relatedness or homology between the sequences is predicted by statistical methods based on weights assigned to the elements aligned between the sequences. We can use two types of methods [2], the tree-based or space-based methods, to compare sequences of amino acids. An important tool in the distance-based mathod is the distance matrix between individual pairs of taxa. In bioinformatics we define the distance between any two nodes in an evolutionary phylogenetic tree using Markov models of residue substitution, such as the Juke-Cantore model for DNA. The necessary condition to make the thory work is that the distance must satisfy metric property. That is, all the entries in the distance matrix representing distances between sequences must satisfy the triangulr inequality, $d(i,k) \leq d(i,j) + d(j,k)$.

2 Sequence-function relationships of protein families

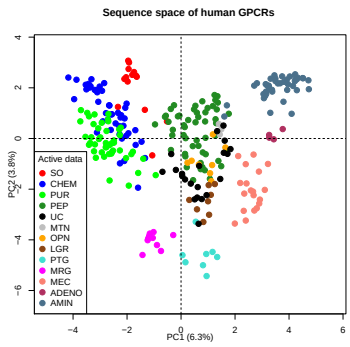
As an example [3], here are presented model data sets gpcr [4] for the human and common fruit fly *drosophila melanogaster* of non-olfactory class A GPCRs. These receptors constitute a large family (283 and 59 sequences, respectively) involved in various types of signal transduction pathways triggered by hormones, odorants, peptides, proteins, and other types of ligands.



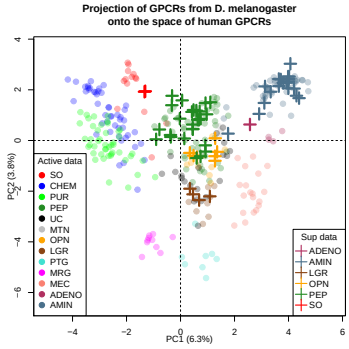
(a) G protein-coupled receptors



(b) Scree plot of metric MDS



(c) Human GPCRs



(d) Projection of GPCRs

Figure 1: MDS analysis (R package bios2md) of the human and common fruit fly *drosophila melanogaster* GPCRs.

Distance matrices are computed from multiple sequence alignments in the FASTA format and can be be visualized in a low dimensional space by metric multidimensional scaling (MDS) (Figure 1). To choose the optimal number of components useful to describe the data in the context of metric MDS the autors used the scree plot 1(b). The aim is to evaluate the number of components required to capture most information contained in the data. The outcome of an MDS analysis is a spatial configuration, in which the objects are represented as points. The points in this spatial representation are arranged in such a way, that their distances correspond to the similarities of the objects: similar object are represented by points that

are close to each other, dissimilar objects by points that are far apart. In figure 1(c) and figure 1(c) the results of MDS analysis are shown in the case of the GPCR sequences from *H. sapiens* and the GPCR sequences from *H. sapiens* with projection of GPCRs from *D. melanogaster*.

3 Bayesian model averaging

Due to the noisy and probabilistic nature of biological signaling data, Bayesian networks have often been considered a good means to learn the causal network behind them. Bayesian Model Averaging (BMA) addresses model uncertainty in a canonical regression problem. If we consider a linear regression model with a constant term, β_0, k explanatory variables $x_1, x_2, ..x_k$ and y being the dependent variable $y = \beta_0 + X\beta_k + \epsilon$, a problem arises when there are many potential explanatory variables in a matrix X. The question is: which variables $X_k \in X$ should be then included in the model, and how important are they? Given the number of regressors, we will have 2^k different combinations of right hand side variables in formula for dependent variable y indexed by M_j for $j = 1; 2; ; ; 2^k$. Once the model space has been constructed, the posterior distribution for any coeficient of interest, say β_h, given the data D is

$$P_r(\beta_h|D) = \sum_{j:\beta_h \in M_j} P_r(\beta_h|M_j)P_r(M_j|D) \tag{1}$$

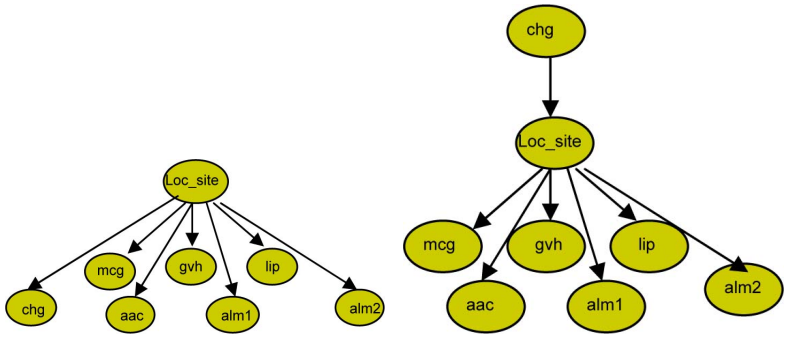
BMA uses each model's posterior probability, $P_r(M_j|D)$, as weights. The posterior model probability of M_j is the ratio of its marginal likelihood to the sum of marginal likelihoods over the entire model space and is given by

$$P_r(M_j|D) = P_r(D|M_j)\frac{P_r(M_j)}{P_r(D)} = P_r(D|M_j)\frac{P_r(M_j)}{\sum_{i:1..2^k} P_r(D|M_i)P_r(M_i)} \tag{2}$$

$$P_r(D|M_j) = \int P_r(D|\beta^j, M_i)P_r(\beta^j|M_i)d\beta^j \tag{3}$$

where β_j is the vector of parameters from model M_j, $P_r(\beta^j|M_i)$ is a prior probability distribution assigned to the parameter model M_j and $P_r(M_j)$ is the prior probability that M_j is the true model.

In this example [6] Bayesian networks is employed to model relationship among seven features (attributes) of E.coli dataset. Attributes [5] which are used are: mcg, gvh, lip, chg, aac, alm1, alm2. Figure 2 ilustrate Bayesian



(a) Bayesian net learned by constrained based approach

(b) Bayesian net learned by score based approach

Figure 2: Bayesian nets learned for the E.coli dataset.

net learned for E.coli dataset by constrained-based 2(a) algorithm (results are same as Naive Bayesian) and score-based 2(b) algorithm.

Acknowledgements

This work is supported by Ministry of Education, Science and Technological Development of the Republic of Serbia.

References

[1] R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and S. Noble, "A statistical framework for genomic data fusion," Bioinformatics, vol. 20, no. 16, pp. 2626–2635, 2004.

[2] Chandrika B-Rao, Kshitish C. Majumdar, Reconstruction of phylogenetic relationships, Journal of Biosciences, Volume 24, Issue 1, pp 121-137, (1999)

[3] Julien Pelé, Jean-Michel Bécu, Hervé Abdi and Marie Chabbert, Bios2mds: an R package for comparing orthologous protein families by metric multidimensional scaling, Bioinformatics 2012, 13:133

[4] Deville J, Rey J and Chabbert M An indel in transmembrane helix 2 helps to trace the molecular evolution of class A G-protein-coupled receptors, J Mol Evol 68, 475- 489 (2009).

[5] Ecoli Data Set, http://archive.ics.uci.edu/ml/datasets/Ecoli

[6] Yetian Chen, Predicting the Cellular Localization Sites of Proteins Using Bayesian Networks and Bayesian Model Averaging, http://www.cs.iastate.edu/ yetianc/cs672/files/CS672Report_YetianChen.pdf

Radiation Effects of Slow Electrons on Biomolecules - Where the Experiment and Theory Meet

Radmila Panajotović [a]
Institute of physics, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia

Abstract

Understanding the effects of high-energy radiation and particles on live cells is one of the most complex and challenging areas of scientific research. Starting from the simple use of X-rays in sterilization of food and water, to the delicate use of the γ-knife in brain surgery, the constant presence of some form of radiation in our natural environment imposes a need to study its effects from many aspects, thus making it truly a multidisciplinary field. Traditionally, the focus in the radiation research, both experimental and theoretical, has been put to the high-energy particles as they carry a lot of energy and can go far through the material. However, a deeper insight into the mechanism of radiation damage inflicted to cells and tissue reveals that the critical damage is caused at the molecular level by the secondary particles among which the most abundant species are low-energy (thermalized)

[a] e-mail address: radmila@ipb.ac.rs

electrons. This paper will give a very short overview of the theory and experiments that consider the interaction of these slow electrons with the DNA and with the DPPC molecules that are usually used as a model for the cell membrane.

1 Introduction

High-energy ionizing radiation, such as γ-rays, X-rays and high-energy particles (neutral and charged), produces a vast number of secondary particles upon they impingement on a material. They are present in Space and on Earth, as part of the regular environmental conditions, as well as a result of human activities in industry and scientific research. Since the early days of discovery of the means to produce high-energy radiation, it has found extensive application in all kinds of technology. As the interest in use of γ- and X-rays grew over time, the danger of exposing humans, animals, and the environment as a whole to the radiation other to the one from natural sources became increasingly significant and important. The truly devastating effects of radiation to life on Earth became ever so evident in the events of nuclear bomb attacks on Hiroshima and Nagasaki, nuclear reactor accidents (Chernobyl, Fukushima, etc.), and nuclear probes in the Pacific. Moreover, the ultimate ambition of the mankind to explore the extra-terrestrial space imposed an additional need to probe, measure and model the effects of high-energy particles (including photons, ions, protons, and electrons) to human crew and plants, animals and food they are taking on board of the Space Stations and spacecrafts. Potential lethality and difficulty to safely confine ionizing radiation in the form of nuclear waste also imposed a strong need for accurate theoretical modelling based on the extensive experimental data, which would describe the physics and chemistry of its interaction with biological material. Finally, research in the field of radiation effects on biological entities is especially beneficial to medicine as it helps to improve the imaging and treatment of tumours and cancer.

2 Sources of high-energy particles

Apart from the X-rays and γ-rays that reach the Earth from distant stars, there are also three kinds of high-energy particles that arrive to our planet from Space: Galactic cosmic rays (GCRs), which are produced by diffusive shock acceleration in supernova remnants and whose flux near Earth is controlled by the solar magnetic activity; Solar magnetic particles (SEPs),

correlated with coronal mass ejections and solar flares and producing ions (>1 MeV per nucleon) and electrons (>100 keV); Radiation belt particles, mainly electrons and protons that are accelerated by the Earth's magnetic field or produced by trapping of SEPs during geomagnetic activity or during CRAND process (Cosmic-ray Albedo Neutron Decay) when the CRs produce neutrons in the planetary atmosphere, which then undergo nuclear reactions and produce electrons and protons inside the geomagnetic field. The experimental measurements are performed by the instrumentation on satellites and by space probes placed in planetary magnetosphere. Due to the shape and strength of the magnetic field, ions are typically of 1 MeV, but electrons can be accelerated from typically 1 keV to 1 MeV ("killer electrons")! The source and mechanism of acceleration have been only discovered in 2007. Fortunately, these particles cannot penetrate through our atmosphere deep enough with such energy, but on their way through, they decelerate and give rise to bremsstrahlung, the emission of the wide spectrum of electromagnetic radiation (X-rays). These X-rays are then mostly absorbed by the molecules in the atmosphere where they then produce a large number of charged particles among which "knock on" electrons (δ rays) are far more energetic than the ions, due to their small mass. This process provides the Earth with the atmospheric layer called ionosphere. Additionally, radioactive elements in rocks and water are also making the overall natural radiation background. Man-made devices, such as particle accelerators, nuclear power plants, medical treatment and security screening devices also contribute to the overall radiation exposure of life on Earth. Therefore, these "knock on" electrons are produced in our environment and all sorts of material, organic or inorganic.

3 Modelling particles stopping power in the material

As we know that the principal process started by the high-energy radiation is ionization, it is clear that the key to successful modelling of the effects of ionizing radiation is to include the effects that secondary particles (electrons ejected from atoms), molecular fragments from dissociation, and induced photon emission produce in the biological material (Figure 1).

This is a very difficult task and solving it starts with the modelling and measurements of the shape and length of the energetic particle path until it thermalizes in the medium. Quantities relevant for measurements and Monte Carlo modelling of these paths are the following: collision stop-

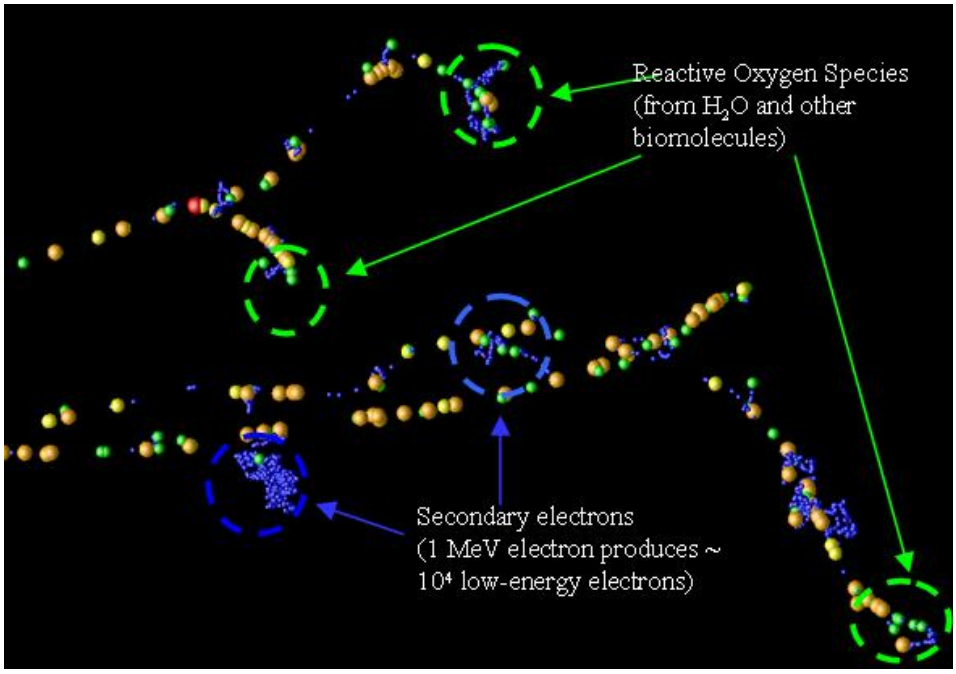


Figure 1: Simulation of the ionizing track of a high-energy electron through a medium. Radiolysis of water creates reactive oxygen species that participate in chemical reactions with other biomolecules.

ping power, as the average rate of energy loss per unit path length, due to Coulomb collisions that result in the ionization and excitation of atoms; density-effect correction, that takes into account the reduction of the collision stopping power due to the polarization of the medium by the incident electron; radiative stopping power that averages the rate of energy loss per unit path length due to collisions with atoms in which bremsstrahlung quanta are emitted; total stopping power presenting the sum of the collision and radiative stopping powers; CSDA (continuous-slowing-down approximation) range - approximation to the average path length travelled by a charged particle as it slows down to rest; projected range equal to the average value of the depth to which a charged particle will penetrate in the course of slowing down to rest; and the detour factor originating from the fact that the multiple scattering makes trajectory of the particle wiggly rather than straight (equal to the ratio of the projected range to the CSDA range (<1)). The main equation used for the modelling scattering of charged particles in the material is Bethe-Bloch formula:

$$-\frac{dE}{dx} = D\frac{Z}{A}z^2\rho\Phi(\beta)(1+\nu) \tag{1}$$

where E is the projectile energy, Z is the atomic number of the element dominant in the material, z is the the elementary charge (in 'e'), ρ is the mass density in g/cm^3, A is the atomic weight of the medium in g/mol, and D is the constant (0.30707 MeVcm^2/mol). Φ factor (interaction probability) is the function of the relative velocity β of the projectile (in units of c):

$$\Phi(\beta) = \frac{1}{\beta^2}\left(\ln\left(\frac{2m_ec^2\gamma^2\beta^2}{I(1+\gamma\frac{m_e}{M})}\right) - \beta^2 - \frac{\delta}{2} - \frac{C}{Z}\right) \tag{2}$$

and depending on the average ionization potential I, mass of the projectile M, and polarisation in the medium and shell screening corrections δ and C, respectively. Factor $(1+\nu)$ is a higher order correction and γ is the velocity factor, $\gamma = \frac{1}{\sqrt{1-\beta^2}}$. The relevant parameters for this formula are contained in the software package GEANT and standalone software libraries that are relevant for calculating detector efficiency in HEP. The ionization energy loss is proportional to the electron density in the medium ($\rho\frac{ZN_A}{A}$) and to the square of the projectile charge. Minimum of ionization (minimum energy loss) is at $\frac{dE}{\rho dx} = 2\ MeVcm^2/g$. For small energies, $\beta \ll 1$, the Bethe formula is reduced to the simpler dependence on ν:

$$-\frac{dE}{dx} \propto \frac{1}{\nu^2}\ln\left(\frac{2m_e\nu^2}{I}\right) \tag{3}$$

Furthermore, the shell screening corrections become much more important as the particle velocity decreases and the electron capture becomes reality. For $\beta \approx 1$ particle, for example, on average only one collision with $E > 1$ keV will occur along a path length of 90 cm of Ar gas [1]. But, for the $\beta\gamma < 0.05$, this formula does not work any longer and only phenomenological fitting formulae are available. For modelling particle track through different materials, including water, DNA, etc. and their stopping power, the data base and software based on Bethe-Bloch equation has been established by NIST [2]. There are three types of particles covered – electrons ("estar"), alpha particles ("astar"), and protons ("pstar"). The energy range covered by this software is from 1 keV up to 10 GeV! The problem with this software is that the uncertainties rapidly rise above 10% for particles with energies below 1 keV. Therefore, the limit for considering the energy of the particles to be low can actually be the energy for which the

velocity of the incident particle is no longer large compared to the velocity of the electrons in the inner shell of the atoms, which is true for electrons that have energies way below 1 keV. The mean of the Bethe-Bloch equation is mostly applied in dosimetry, where the energy is deposited in the bulk. In biological targets the "bulk" presents a mixture of a vast number of functional tissue, with different density and molecular structure, as well as different physiological connections within the entire body. Therefore, experiments that would produce conclusive results which accurately describe the effects of radiation on human or animal organs are extremely limited and difficult, which often reflects in either a relatively small statistics or a very long and not very rigorous medical research on the existing victims of radiation from nuclear accidents, for example.

At low energies, electrons primarily lose their energy by ionization, without loosing their energy by emitting the photons. The ionization energy loss rises logarithmically with the drop in electron energy. Now the question is how this value depends on the type of the target material. A mixture or compound can be thought of as made up of thin layers of pure elements in the proportion defined by the target in question (Bragg additivity). The problem here is the averaging of the ionization potential, because the electrons in a compound are bound differently than in free elements; hence the electron density becomes the key factor in calculations. Having the energy that keeps them around ions and molecules in the medium for longer, low-energy electrons can also recombine with the ions and molecular fragments. The number of products along the ionization track (Figure 1) is consequently very high and not fixed in space and time, which imposes the need to know probabilities for each of the processes involved as accurately as possible. These probabilities are then used for Monte Carlo simulations and calculations of radiation dose for radiation therapy and radiation protection or for the estimates of the risks of exposure of the astronauts in spacecrafts and space stations, for example.

4 Why are low-energy electrons dangerous to cells?

So, why is the radiation damaging to life on Earth and which elements in living organisms are affected the most? To answer this question, we need not only to study the development of the disease and damage to the organs in the body, but also to understand the molecular structure of the cell and the physics and chemistry (Figure 2) behind the interaction of radiation with a large number of biologically relevant molecules (DNA, proteins, lipids, water, etc.).

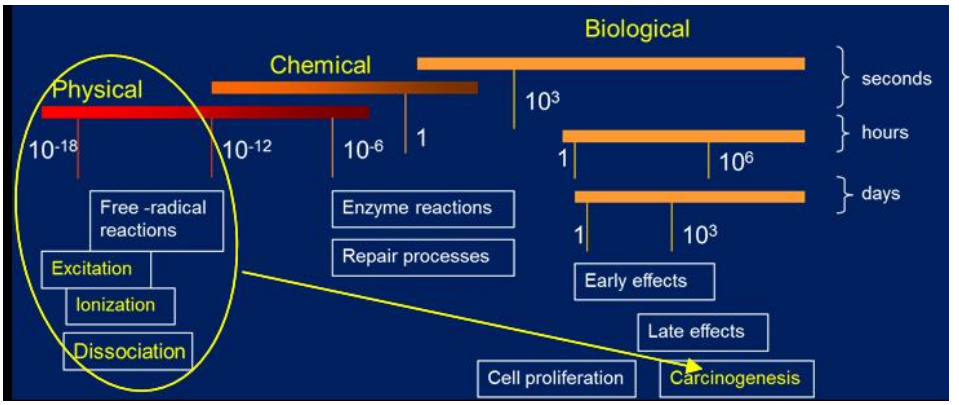


Figure 2: Time scale for effects of radiation at the different level of complexity; electrons are involved in the processes of excitation, ionization and dissociation of molecules within the cell.

Biological effects of radiation are reflected through the deterministic effects (short − term large dose exposure causing fast death) and stochastic effects (random radiation-induced changes in the cells at the molecular level that lead to tumours and cancer). The main parameter in this assessment is the absorbed dose, i.e. energy deposited in the body, and the effective dose, which is the weighted energy deposited from different types of radiation in different organ tissue. The main data for modelling this kind of biological effects are collected, unfortunately, from the victims of nuclear bomb explosions and accidents. A major part of the large body of work dedicated to the effects of radiation on living organisms is focused around the damage that charged particles (ions and electrons) inflict to the DNA molecule. This damage can be direct (from ionizing radiation) and indirect (from radiolytic decomposition of water) in the form of short-term and long-term effects (accumulation of damage). The most important are the single-strand and double- strand breaks, clustered damage on nucleic bases, etc. The cell's response to these changes depends on its role in the organ, i.e. type of tissue, and on whether it is healthy or cancerous. It can lead to cell's death (apoptosis), creation of mutations that lead to cancerogenesis, or the cells may remain resistant to radiation owing to their efficient repair mechanism or another (mostly unknown!) biochemical mechanism of protection. In general, cells and tissue that grows fast (blood-forming organs, gastrointestinal tract, hair follicles, etc.) exhibit the highest sensitivity to radiation, while the opposite is true for neurons in brain and the muscular tissue. In order to understand what is the un-

derlying physicochemical mechanism of the DNA damage, it is necessary to understand what is the destiny of the secondary electrons with very low energy (< 30 eV), because of the fact that they are the most abundant secondary species in the process described. It turns out that these electrons induce fragmentation of molecules, thus creating potentially reactive ions and oxygen species [4], which show to be extremely damaging to the DNA (and not only to DNA!) [5]. Investigation, both experimental and theoretical, of the vibrational and electronic excitation of nucleic bases and nucleosides by low-energy electron impact leading to dissociation through formation of temporary negative ions aims at understanding the correlation of nucleic base-sugar moiety conformational coupling and its consequences on the bond cleavage in DNA. For example, the conformation of the 2'-deoxyribose moiety (Figure 3) with respect to the base is expected to influence which species are formed upon exposure of nucleosides to ionizing radiation. Furthermore, the holes created in the place of ionized moieties move around the affected site along with electrons and protons. These processes are fast and extend over several nucleic bases in the DNA chain.

Generally speaking, they all depend on the sequence of the bases in the DNA molecule, their electron affinities and the hydration level at different sites. The high level *ab initio* theories such as Density Functional Theory (DFT) have been used to describe the mechanism of these processes with significant success [6]. One of the contributors to the overall mechanism of the nucleic base damage and strand break is the dissociative electron attachment (DEA) in which a transient negative ion is formed from the molecular moiety which temporarily captures the incoming electron of very low energy (below the dissociation limit, typically, $E < 15$ eV). Electrons with these "thermal" energies are captured to the $\pi^\star$ states of the nucleic bases with the largest electron affinity, i.e. thymine and cytosine. The holes (cation radicals), as a consequence, migrate to the bases with lowest ionization energy, which is guanine. On the other hand, holes from the sugar or phosphate moieties may have different destiny - apart from transferring to the base, they may incite deprotonation of the sugar cation radical which then becomes neutral. This process may be critical for strand breaks in DNA, which are considered to be the most important DNA damage. This is even more important when the damage is clustered around the same site and produces double strand breaks that are difficult for the cell to repair successfully. In recent years, as the computational resources grew and became more widely available, the theoretical modelling of many-body processes from first principles became possible on a more realistic time scale (days

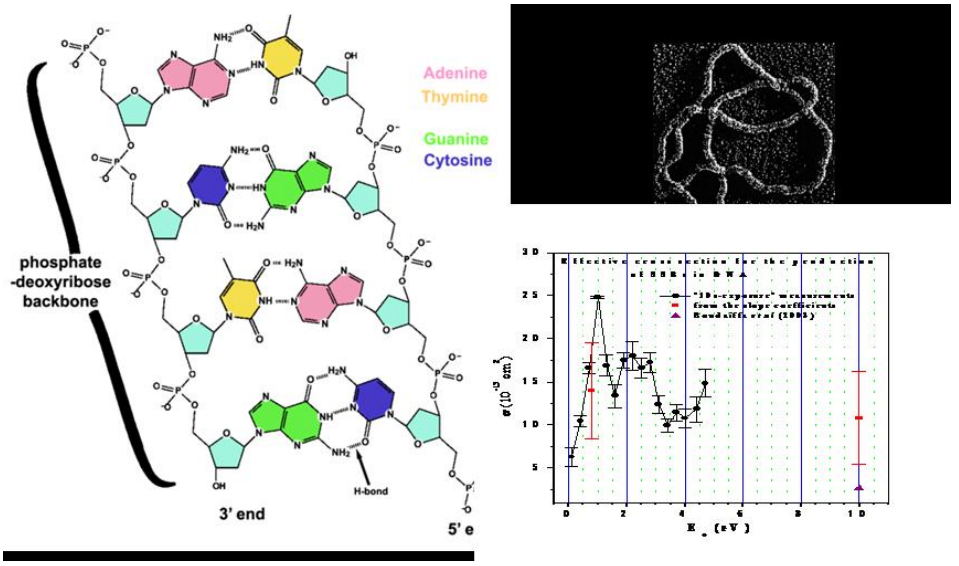


Figure 3: Schematic structure of the DNA molecule with four nucleic bases connected with the sugar moiety and linked together by means of the phosphate bond (left). The experiment [3] with low-energy electrons impinging on the supercoiled plasmid DNA (top-right) from *Escherichia Coli* showed that the cross section for single strand breaks increases significantly for incident electron energies in the range from a few hundreds of eV up to almost 4 eV (the plot on the right). This feature in the excitation function is an evidence of a resonance that enhances the probability for a strand break 2.5 times.

instead of months). Usually, the approach is through using sophisticated *ab initio* Hartree-Fock (HF), Moeller-Plesset perturbation theory (MP2), and the DFT. The basic premise is that the interaction of the incoming electron with N electrons in the molecule is a many-body problem that needs to be reduced to a simpler one by means of different approximations. One of them is, for example, the so-called static exchange approximation [7], where polarization-correlation potential is used. The exchange potential is described through the local-density approximation which is applied to many other molecules and gives a good qualitative picture of the formation of temporary negative ions. In the case of DNA, the large dipole moment of nucleic bases requires that the long-range effect of the dipole field on the scattered electron has to be taken in to account. This fact makes the calculation very time consuming, so the dipole potential and centrifugal po-

tential are combined to match the typical dipole moment of nucleic bases ($2 - 2.5$ D) and keep the number of dipole-bound states finite. Typical overestimates for cross sections and resonance energies obtained in this model are from $1 - 2$ eV with respect to those measured in experiments. Considering experiments in which pure, dry DNA and its building blocks are exposed to a low-energy electron beam [1], they are usually performed in very controlled conditions – ultra-high vacuum – in order to provide conclusivness and reproducibility of the data. Mechanisms that can lead to specific features, such as the strong peak at 1 eV in (Figure 1), are not easy to clearly identify, but there are several factors that can be identified as playing a more or less important role in the SSB damage of DNA. At energies below 5 eV, the possibility of secondary electrons coming from ionization of the sample by the primary electron beam can be excluded because the ionization of any of DNA constituents requires at least 5.7 eV (for the fully solvated GC pair and more for the unsolvated) up to 11eV [8] (sugar-phosphate backbone fragments), and the formation of OH radicals from water molecules requires higher energy for direct dissociation; no dissociative electron attachment (DEA) signal arising from structural water has been observed from vacuum-dried DNA films [8] either. It has been shown [9, 10] that the DNA backbone lesions caused by sub-excitation energy electrons are essentially the result of breaking of the C-O bond between the sugar and a phosphate moiety. At energies below 3 eV, low-energy electron can attach to a $\pi^\star$ orbital of the phosphate group forming a TNI or, the resonance capture of the lowest $\pi^\star$ orbital of the bases followed by the electron transport to the $\pi^\star$ of the phosphate. Both mechanisms would lead to the transition state formed with an extra electron in the usually unfilled P=O $\pi^\star$ orbital that can further lead, via curve-crossing, to a $\sigma^\star$ anion state, and to cleavage of the phosphodiester bond. Since electronic excitation of the nucleic bases is not possible at such low energy (the lowest-lying triplet electronic state in thymine is around $3.7 - 4$ eV), the transient anion is produced via a shape resonance, unless vibrational Feshbach resonances are involved. Recently, instead of using a traditional preparation of thin films of DNA or nucleic bases for the HREELS (High Resolution Electron Energy Loss) measurements, where the integral signal over all molecules in the sample is detected, a very advanced technique of DNA origami using the molecular manipulation by the Atomic Force Microscope has been successfully applied to observe the strand breaks in DNA [11]. A drawback to this approach is, of course, the artificiality of conditions compared to the ones in real life, but the advantage is that the understanding of the underlying principles of the DNA damage is easier to achieve.

5 Beyond DNA damage

Apart from the DNA and its constituents, owing to the vast number of metabolic processes that can lead to potentially damaging radiation effects, the cell membrane with its highly ordered structure and complex functionality relying on the activity of various proteins, presents another important target for cancer research (for example, a fully saturated phospholipid, DPPC, is a major component of the highly radiation-sensitive lung tissue) and a perfect "live model" of a biosensor. Radiation damage of molecules that compose the cell membrane, such as phospholipids, proteins and polysaccharides, can be crucial for the operation of cell's ion transport and signaling. It is a challenging task for both experiment [12] and theory [13], mostly due to the difficulty in recreating a reproducible, stable and "true-to-life" sample for various experimental techniques (KPFM, STM, FT-IR, XPS, NEXAFS, HRLEES, etc.) to be applied and, in the case of theoretical modeling, properly including the forces that govern the interaction between the individual components in a bio-molecular complex while keeping the computation time relatively short. When the DPPC molecules were deposited [12] as a thin film on a gold-coated silicon substrate or on a silicon wafer and irradiated by electrons of energy between 5 and 200 eV, the shifts and intensity of the binding energies of C 1s, O 1s, P 2p, and N 1s atoms were observed through analysis of the photo-electrons emitted from the target before and after electron irradiation. Overall, major damage to the monolayer film was caused by cutting the methyl groups from nitrogen and phosphate group from the rest of the molecule. The least effect of electron irradiation is shown on the P 2p band, regardless of the incident energy. The effects are significantly smaller for 5 and 200 eV electrons than for energies in between, which is indicating that the damage to the real-life cell membrane could in fact originate from the damage of the numerous other molecules embedded in the lipid matrix. Besides the well known DFT [13] theory applied to the analysis of the radiation damage to polymers, in the case of large molecules, the Molecular Dynamics software packages have been extensively used to describe the self-assembly and conformational changes in proteins, lipids and nucleic acids. One of them, and very popular for its speed, is GROMACS software package [14] that can run on the CPUs and GPUs alike. It is an "open source" software that is continually being updated and is being adapted for parallel computing. So far it has not been used for theoretical modelling of radiation damage, but this may be due to the already proven success of the existing DFT codes.

6 Conclusion

The study of the effects of radiation on biomolecules and especially of the low-energy electron interaction with DNA, proteins, lipids, and other molecular species relevant for life on Earth is going to remain an active field of research for many decades. Theoretical models in physics and chemistry, as well as the experimental and clinical research in medicine and biology, are being developed side to side, benefitting from the new advances in computer technology and software development. This development will be of great value not only for the purpose of pure science, but also for the protection and prosperity of our environment and life on Earth.

Acknowledgements

This work was supported by the Ministry of education, science and technology of the Republic of Serbia, grant number OI171005.

References

[1] H. Bichsel, Nucl. Instrum. Methods A562 (2006) 154–197

[2] NISTIR 4999, Stopping Power and Range Tables, www.physics.nist.gov/PhysRefData/Star/Text/programs.html

[3] R Panajotovic, F Martin, P Cloutier, D Hunting, and L Sanche, "Effective Cross Sections for Single Strand Break Production in Plasmid DNA by to 4.7 eV electrons", Radiation Research, 165 (2006) 452-459

[4] Radmila Panajotovic and Leon Sanche, "From DNA to nucleic bases - the effects of low-energy electron impact", Journal of Physics: Conference Series, 88 (2007) 012074

[5] R Panajotovic, M Michaud and L Sanche, "Cross sections for low-energy electron scattering from adenine in the condensed phase", Phys. Chem. Chem. Phys. 9 (2007) 138

[6] M. K. Shukla, J. Leszczynsky (eds.), Radiation induced Molecular Phenomena in Nucleic Acids (2008) 577-617, Springer Science+Business Media B. V.

[7] S. Tonzani and C. H. Greene, J. Chem. Phys. 124 (2006) 054312

[8] Colson A O, Besler B and Sevilla M 1993 J. Phys. Chem. 97, 13852-13859 and 8092; Morgan L A 1998 J. Phys. B 21, 5003-5011;

[9] Abdoul-Carime H and Sanche L 2001 Radiat. Res. 156, 151-157; Aflatooni K, Gallup G A and Burrow P D 1998 J. Phys. Chem. A, 102, 6205-6207

[10] Barrios R, Skurski P and Simons J 2002 J. Phys. Chem. 106, 7991-7994; Li X, Sevilla M and Sanche L 2003 J. Am. Chem. Soc. 125, 13668-13669

[11] Adrian Keller, Ilko Bald, Alexandru Rotaru, Emilie Cauët , Kurt V. Gothelf, and Flemming Besenbacher Probing Electron-Induced Bond Cleavage at the Single-Molecule Level Using DNA Origami Templates, ACS Nano, 6 (5) (2012) 4392–4399

[12] Panajotovic, R, Schnietz M, Turchanin A, Mason N and Gölzhäuser A, "XPS Study on Effects of Electron- Beam Irradiation of Thin Condensed DPPC Films", Book of Abstracts, ECAMP X, 4.-10. June 2010, Salamanca, Spain

[13] S. L. Dudarev, "Density Functional Theory Models for Radiation Damage", Annual Review of Materials Research, 43 (2013) 35-61

[14] http://www.gromacs.org/

An Integrative Approach to Relating Genotype, Phenotype and Taxonomic Characteristics in Prokaryotes – An Overview

Gordana Pavlović-Lažetić [a]
Faculty of Mathematics, University of Belgrade, Serbia
Vesna Pajić [b]
Faculty of Agriculture,University of Belgrade, Belgrade, Serbia
Nenad Mitić [c]
Faculty of Mathematics, University of Belgrade, Serbia
Jovana Kovačević [d]
Faculty of Mathematics, University of Belgrade, Serbia
Miloš Beljanski [e]
Institute of General and Physical Chemistry, Serbia

Abstract

Correlations between specific genotype, phenotype and taxonomic organism characteristics, such as genome size, genome GC content, proteome size and distribution among functional groups of

[a] e-mail address: gordana@matf.bg.ac.yu
[b] e-mail address: svesna@agrif.bg.ac.rs
[c] e-mail address: nenad@matf.bg.ac.rs
[d] e-mail address: jovana@matf.bg.ac.yu
[e] e-mail address: mbel@matf.bg.ac.yu

proteins, optimal growth temperature, habitat, oxygen requirements, for different superkingdoms, phyla, or even species, have been the subject of many studies with different outcomes.

As opposed to genotype and taxonomic data which is usually well structured and deposited into databases, data on phenotypic characteristics of organisms are often found in scientific papers or encyclopedias, in an unstructured or semistructured form. We use the finite state method for literature mining and its application to the Encyclopedia of Prokaryotes in order to integrate the data obtained from literature with the most comprehensive formatted database - NCBI Entrez Genome database. Using a larger dataset of prokaryotes of different taxons (i.e., superkingdoms and phyla), we reconsider high-expectation correlations between ge-nomic, phenotype and taxonomic characteristics. Further, we apply algorithms for association rule mining in order to identify the most confident associations between specific modalities of the characteristics considered. The results of both literature mining and correlation-association rules mining are presented and commented on.

1 Introduction

One of the main problems in the post-genomic era is to correlate or associate phenotypic characteristics of an organism to composition, genes and proteins encoded by its genome, for as large and diverse a collection of organisms as possible. Such relationships can be best analyzed using mathematical and computational methods and techniques.

Publicly available databases, such as the one maintained by the National Center for Biotechnology Information, NCBI [1] provides information for a variety of prokaryotic genomes (both superkingdoms Archaea and Bacteria). Still, much larger portion of such information resides in semi-structured and unstructured forms such as encyclopedia, articles, books, web pages and other document and literature sources. These kinds of resources need to be transformed into structured forms (e.g. databases), in order to be processed and utilized.

The overall research goal in the field is to analyze relations between genomic (genotypic) features and their phenotypic characteristics (metabolic and physiological characteristics, morphology, lifestyle and habitat adaptations, etc), for different taxonomic categories – superkingdoms bacteria and archaea, specific phyla and species. These relationships provide for deeper comprehension of evolutionary processes and for some prediction possibilities, e.g., trends prediction of some pandemia.

The main goal of the work presented is twofold. First, it is aimed at systematically and multiply comparing genomic, phenotypic and taxonomic characteristics of prokaryotic organisms from a data collection as comprehensive as possible, and at observing their cross correlations and associations. Second, it is to present and illustrate how a methodology for information extraction from semi-structured sources and their integration with structured databases into a common framework, can contribute to comprehensiveness of the data collection as a resource.

2 Related work

There are quite a research on relating genotypic, phenotypic and taxonomic characteristics of organisms, especially prokaryotes [2-14].

The fact that microbial phenotypes are typically due to the joined action of multiple gene functions is emphasized in [2]. Jim et al. consider inferring gene function from cross-organismal distribution of phenotypic traits, which is a reliable approach when the phenotype does not arise from many alternate mechanisms. Burra et al, [3] analyze thermal adaptation vs. structural disorder and suggest that adaptation to extreme conditions is achieved by a significant functional simplification; in [4], Goh et al. present a systematic approach to discovering genotype-phenotype associations that combines

phenotypic information from a biomedical informatics database, GIDEON, with the molecular information contained in National Center for Biotechnology Information's Clusters of Orthologous Groups database (NCBI COGs); in [5], the distribution of G + C content against chromosome size among 640 fully sequenced bacterial chromosomes is analyzed, suggesting that the shorter chromosomes have a wide, 20–70%, range of G + C contents while the longest chromosomes are restricted to near 70% G + C content; in [6], an in-depth analysis was conducted to evaluate various potential intrinsic and extrinsic factors in association with GC content variation among eubacterial genomes; in [7], an attempt has been made to explain why GC content of bacterial genomes varies from 25 to 75%; it is argued that genomes of bacteria that rely on their host for survival (obligatory pathogens or symbionts) tend to be AT rich; in [8] the authors find that with increasing habitat temperature and decreasing genome size, the proportion of genomic DNA in intergenic regions as well as generation time decreases; in [9], quantity GCVAR is introduced, the intra-genomic GC content variability with respect to the average GC content of the total genome; GCVAR is found to be significantly higher among anaerobes than both aerobic and facultative microbes, and that it varies greatly among phyla. In [10] the results are presented giving further support to the link between aerobic respiration and genomic GC content. In [11], the influence of lifestyle of prokaryotes on the correlation between genome size and genomic GC content has been analyzed - the effect of optimal growth temperature, aerobics / anaerobics, motility, aquatics and parasitism, using data from 411 representative prokaryotic species (including archaea and bacteria) and correlation/regression approaches; in [12], results are reported on a comparative analysis of GC composition and optimal growth temperature for over 100 prokaryotes. Musto et al. 2006 [13] demonstrated, for some bacteria families, that there exists relationship between genome size and GC level for aerobic, facultative, and microaerophilic species, but not for anaerobic prokaryotes. Furthermore, Mann and Chen [14] found that larger genomes (more than 3 Mb) in free-living organisms, as a result of more complex and varied environments, show trend toward higher GC content than smaller ones, while nutrient limiting and nutrient poor environments dictate smaller genomes of low GC.

2.1 Association rule mining

In bioinformatics, association rule mining has been used primarily in microarray and gene expression data analysis (Creighton and Hanash [15], Georgii et al. [16], Carmona-Saez et al. [17], Martinez et al. [18], Martinez et al. [19], Gyenesei et al. [20]).

Tamura, D'haeseleer [21] developed an association rule mining algorithm NETCAR for extracting sets of COGs (clusters of orthologous groups of proteins) associated with a phenotype from COG phylogenetic profiles and a phenotype profile.

MacDonald & Beiko [22] have developed a new genotype–phenotype association approach that uses Classification based on Predictive Association Rules (CPAR).

2.2 Information extraction

Some of the studies of relating and associating different organism characteristics include the use of literature mining and information extraction from different text documents, such as scientific papers or encyclopedias. Korbel et al, [23] use literature mining and comparative genome analysis for association of genes to phenotypes; Jimeno-Yepes et al [24] exploit ontological resources for searching genes and related diseases from scientific literature. In Pajic et al [25, 26] a two-phased method for information extraction from semi-structured resources is presented and applied to extracting information from an encyclopedia of bacteriology, as a data source complementing the existing formatted databases and an extended material for relating organism characteristics. The method and the resulting database will be briefly presented in section 4.2. At last (but not least), in [27] we presented some results on associating characteristics from the extended resource.

3 Materials

There are plenty of genotype data and gene sequences for different organisms, usually well structured and residing in databases; still, data on phenotypic characteristics of organisms can be often found scattered across different text documents, e.g., scientific papers or encyclopedias.

Some of the most comprehensive public databases with all kinds of information – genotypic, phenotypic and taxonomic, are, for example, NCBI Entrez Genome database (the most extensive) [1], Comprehensive Microbial Resource [28], Genome Atlas Database [29], IMG [30], PATRIC databases [31], databases and tools for specific types of genotype-phenotype research, etc.

Two data sources have been used so far in our research:
(i) NCBI Entrez Genome database [1] - an instance from 2011 (table *Organism_info)*, and
(ii) *Bergey's Manual of Systematic Bacteriology* [32-34]

3.3 NCBI Entrez Genome Database: Table *Organism_info*

The characteristics present in the NCBI database include genotype characteristics, such as genome size and GC content, phenotypic characteristics such as shape, oxygen, habitat, salinity, temperature, gram stain, motility, pathogenicity and taxonomic characteristics such as superkingdom, phyla, species. While the genotype is genetic constitution of an organism, phenotype refers to any observable characteristics or organism trait, such as its morphology, development, biochemical or physiological properties, or behavior. Phenotypes result from the expression of an organism's genes, as well as the influence

and interactions of environmental factors. Some of the considered characteristics are the following.

Genome size is the total amount of DNA contained within one copy of a genome. It is measured as the total number of nucleotide bases pairs. In known prokaryotic organisms genome size vary, for example, between Candidatus Carsonella ruddii (an obligate endosymbiotic Gammaproteobacteria) with a genome of 160 000 bp (Nakabachi et al. [35]), to 10,148,695 bp for Streptomyces scabiei 87.22 (an important bacterial plant pathogen). Distribution of genome size in prokaryotes, calculated by Koonin and Wolf [36], clearly separates two broad genome classes with 4Mb border.

Guanine-Cytosine (GC) content (or ratio) of a genome refers to the percentage (or ratio) of nitrogenous bases of genome nucleic acids. It may vary between the genomes, as well as in the genome. Due to the nature of the genetic code it is virtually impossible for an organism to have a genome with a GC-content of either 0% or 100% (Gardiner [37]). Average GC content of bacterial genomes varies in range from 25% to 75% (Mann and Chen [38]).

Habitat. Bacteria grow in a wide variety of habitats and conditions. They may be found on the highest mountains, the bottom of the deepest oceans, in the animals guts, and even in the rocks and ice (Schlegel and Jannasch [39]). Modalities for habitat, found in the NCBI database [1], are aquatic, multiple, specialized (i.e., hot springs, salty lakes), host-associated (i.e., symbiotic) and terrestrial.

Oxygen requirement. Bacteria have a wide range of environmental and nutritive requirements. Most bacteria may be placed into one of four groups based on their response to gaseous oxygen. Modalities for oxygen requirement found in [1] are aerobic, facultative anaerobic (facultative for short), anaerobic and microaerophilic. Aerobic bacteria grow in the presence of oxygen and use it as a terminal acceptor of electrons in respiratory chain. Microaerophilic bacteria require lower level of oxygen than present in atmosphere. Anaerobic bacteria instead of oxygen use some other inorganic electron acceptor (sulfur, for example). Facultative anaerobe use oxygen when present, but may grow without oxygen. As compared to anaerobic, aerobic prokaryotes have shown increased GC content (Schlegel and Jannasch [39]).

Temperature range. Bacteria grow in many environments from arctic oceans to hot springs. They can be classified into the following modalities: mesophile and extremophile, i.e., thermophile, hyperthermophile and cryophile (or psychrophile). A mesophile grows best in moderate temperature, between 15ºC and 40ºC. The habitats of these organisms include soil, human or animal body, etc. Thermophiles are extremophilic organisms that prefer relatively higher temperatures, between 45ºC and 80ºC. Many of them belong to Archaea. Hyperthermophiles are extreme thermophiles which prefer temperatures above 60ºC. Psychrophiles or Cryophiles are extremophilic organisms that are capable of growth in cold temperatures below 15ºC and are common in cold soils, polar ice or cold waters.

In the NCBI table Organism_info there were 2163 different species with 7467 isolates. The table was under half-populated. Fig. 1. Presents partial statistics on the table Organism_info (number of isolates / species with data defined for oxygen requirement, habitat, temperature range, gram stain and their combinations).

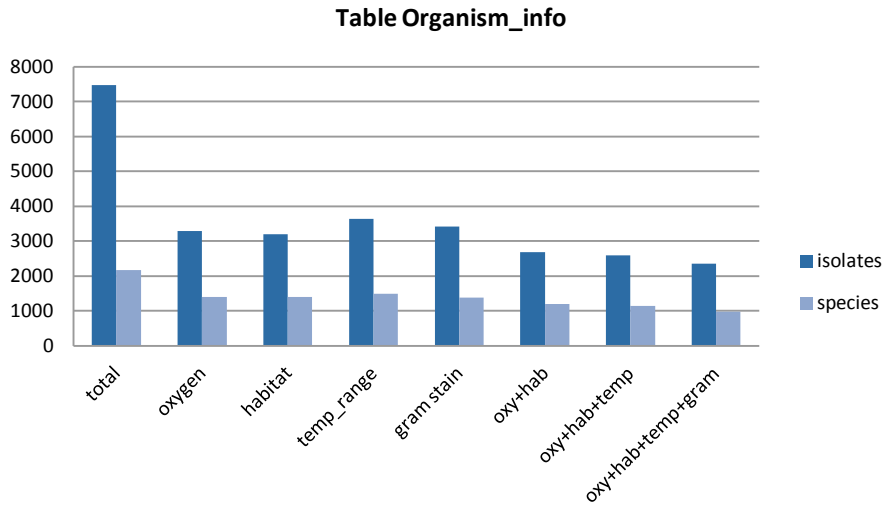


Fig. 1. Statistics on the table *Organism_info* (NCBI)

Distribution of modalities for oxygen requirement, habitat, temperature range and gram stain is presented in Table 1.

OXYGEN					
Undef	Anaerobic	Facultative	Microaerophilic	Aerobic	
4182	792	1300	129	1064	
HABITAT					
Undef	Host-associated	Specialized	Multiple	Aquatic	Terrestrial
4273	1429	204	856	502	203
TEMP_RANGE					
Undef	Psychrophilic	Cryophilic	Mesophilic	Thermophilic	Hyperthermophilic
3835	35	1	3377	141	78
GRAM STAIN					
Undef	+	-	_		
4043	1371	2047	6		

Table 1. Distribution of modalities for phenotypic characteristics (NCBI Organism_info)

3.4 Encyclopedia of Microorganisms

As a literature source we used the encyclopedia of microorganisms - *Bergey's Manual of Systematic Bacteriology, Volume 2 : The Proteobacteria* [32], *Bergey's Manual of Systematic Bacteriology, Volume 3: The Firmicutes* [33] *and Bergey's Manual of Systematic Bacteriology, Volume 4: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes* [34]. Its structure, which is of interest for the information extraction method, is described in [25 - 27]. It contains the taxonomic hierarchy and then description of specific taxons (e.g., genera, species). Descriptions are unstructured or semistructured, and an example of species description is presented in Fig. 2 (underlined are values of characteristics to be extracted).

9. **Hyphomicrobium zavarzinii** Hirsch 1989b, 495[VP] (Effective publication: Hirsch 1989, 1903.)
za.var.zin' i.i. M.L. gen. n. *zavarzinii* of Zavarzin, named for G.A. Zavarzin, the Russian microbiologist who isolated these bacteria.

Mother cells drop- or pear-shaped, somewhat slender, with hyphae that rarely branch. Mother cells 0.63 × 1.8 µm (range: 0.5–0.9 × 0.7–2.5 µm). Swarmer cells with 1–3 subpolar flagella. In liquid media under most growth conditions, rosettes are formed, since mother cells produce a polar holdfast. Growth in liquids initially as turbidity and later as a pellicle, with precipitation on the bottom. Colonies on solid media are colorless to light brownish or beige, smooth and shiny, with entire edges.

Chemoorganotrophic, aerobic, oligocarbophilic. Good growth with the following carbon sources: methanol, methylamine·HCl, formate, *n*-butyrate, isovalerate, crotonate, β-hydroxybutyrate, ethanol, *n*-propanol, isobutanol, and glycerol. Growth is stimulated significantly by acetate, *n*-valerate, α-oxoglutarate, galacturonate, formaldehyde, D-glucose, D-mannose, D-melibiose, amygdalin, esculin, chitin, Bacto peptone, DL-lysine, DL-aspartate, and dilute human urine. Nitrogen sources utilized are: NH_4^+, NO_2^-, NO_3^-, and (poorly) Bacto peptone. There is slow growth in the absence of added nitrogen sources (oligonitrophily). Poor growth on sheep blood agar with α-hemolysis. The following antibiotics inhibit growth at 30 µg (per disc): kanamycin, neomycin, and tetracycline. Streptomycin at 10 µg is also inhibitory. There is growth in the presence of 3.5% NaCl. Temperature range: 15–37°C. Optimal pH: 6.5–7.5. Visible light inhibits growth slightly.

Grow anaerobically with nitrate and gas formation (with methanol as the carbon source). With methylamine·HCl and thioglycolate, there is little growth. Catalase and cytochrome oxidase are positive; gelatin liquefaction is negative. Poly-β-hydroxybutyrate is a storage product.

Not pathogenic for mice or guinea pigs.

Genome size: 2.73 × 10^9 Da (strain ZV-580; Kölbel-Boelke et al., 1985).

Habitat: peaty and moist soil near Moscow, Russia.

The mol% G + C of the DNA is: 61.8–64.8 (Bd, T_m, HPLC) (Mandel et al., 1972; Gebers et al., 1986; Urakami and Komagata, 1987b; Urakami et al., 1995b).

Type strain: ATCC 27496, IFAM ZV-622.

GenBank accession number (16S rRNA): Y14305.

Additional Remarks: Additional strains include IFAM ZV-580, ZV-620, MY-619, MC-625, MC-629, MC-630, and MC-627.

Fig. 2. Description of the species Hyphomicrobium zavrziniie from the encyclopedia 'Systematic Bacteriology'; the underlined parts of the text represent values to be extracted

4 Methods

Methodology underneath the research presented involves methods and tools for correlation analysis, association rule mining and text mining – information extraction. Different pairs of genomic and phenotypic characteristics are differently correlated for different taxonomic categories (or other phenotypic characteristics), e.g., genome size and optimal growth temperature may be negatively correlated for bacteria and uncorrelated for archaea). Correlation analysis gives valuable insights into potential interdependencies among these characteristics.

4.1 Association rule mining

Association rule mining proves useful in determining significant co-occurrences of specific characteristics. The task is formulated as follows:

*Given a set of transactions consisting of one or more elements (*items*), find rules that predict occurrence of an item based on occurrence of other items in the transaction.*

Association rules are of the form A→B where A and B are sets of elements represented in the data set. A is called *body* of the rule, and B - *head* of the rule. Implication refers to co-occurrence, not to causality.

There are several measures for quality estimation of the rules discovered. The most often used are *support* and *confidence*.

Support for the rule A→B, denoted by $s(A \rightarrow B)$, is defined as

$$s(A \Rightarrow B) = \frac{\sigma(A \cup B)}{N}$$

where $\sigma(X)$ denotes number of occurrences of an item X in a transaction, and N is total number of items.

Confidence measures how often item B occurs in transactions containing item A, and for the rule A→B, it is defined as

$$c(A \Rightarrow B) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

The goal of mining association rules is finding all the rules, for a transaction set T, with support ≥ minsup and confidence ≥minconf.

We applied algorithms for association rule mining from the data mining system, IBM Intelligent Miner. It is a part of the programming package IBM InfoSphere Warehouse V9.5 (and later versions) (http://www-01.ibm.com/software/data/infosphere/warehouse/mining.html). In this implementation association rule mining is based on the *Apriori* algorithm and fast insight into the discovered rules can be obtained by a sophisticated visualization component.

4.2 Information Extraction

Since much more information about microbes and other organisms can be found in unstructured and semi-structured documents, we applied a text mining technique – information extraction with the primary goal of mining genotype / phenotype organism characteristics from text.

Information extraction is a process of finding specific data form unstructured texts, in order to use them or store them into a structured database. Among two broadly present approaches to information extraction – statistical and rule-based, we follow the rule-based one. We use the two phase method based on finite state transducers (FST) for information extraction from text, introduced in [39]. Every attribute, that is, every property or characteristic of the organism, which needs to be extracted from the text, is recognized and outputted by a particular FST we created. It extracts the relevant data from text segments by applying a collection of FSTs in the form of graphs, describing possible ways the information we are interested in can be expressed in the text (data to be extracted). As a tool for dealing with FSTs we used the system UNITEX [40]. We developed the FSTs for each of the characteristics considered. For example, the FSTs for gram stain is represented in Fig. 3 and will recognize the following text sequences and extract the corresponding data “pos” or “neg”:

"Gram-positive"
"gram-pos"
"gram pos."
"Gram + "
"gram negative"
"Gram-negative"

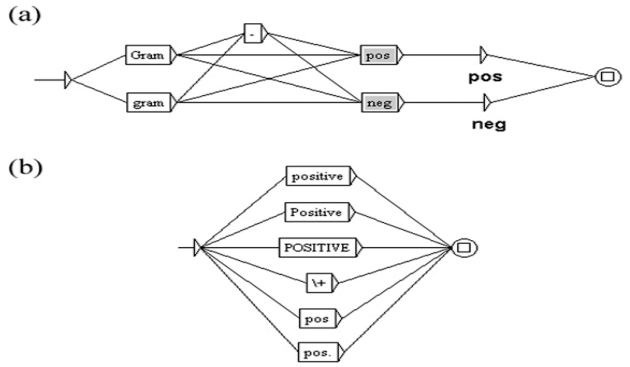


Fig. 3. Transducer for extracting information about the genome gram stain created with UNITEX: (a) The transducer contains calls to sub-graphs *pos* (positive) and *neg* (negative) and outputs the corresponding mark (pos / neg); (b) The *pos* sub-graph for describing ways of specifying the positive value; similar sub-graph exists for negative gram stain.

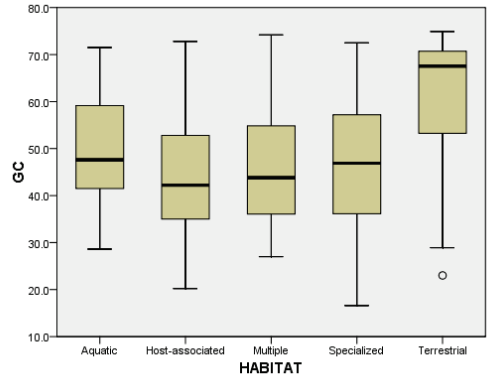
The extracted data was then put into the database which was used for further analysis.

5 Results and discussion

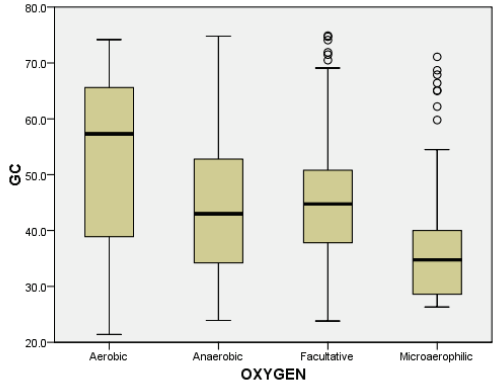
5.1 Correlations analysis from the original NCBI data

We reconsider correlations among different characteristics of prokaryotes and extend the study to multiple characteristics correlations. Characteristics considered are genome size, GC content, habitat, oxygen requirement and temperature range. For genome size, we consider "low" and "high" genome size, with 4Mb border, as explained before. For GC content, we consider three modalities: (less than mean value - standard deviation), medium (in the interval mean value ± standard deviation), high (greater than mean value + standard deviation). Modalities for habitat, oxygen requirement and temperature range have been described in Introduction. Some of the results show that:

1. GC content:
    a. most bacteria have medium GC content;
    b. terrestrial and aerobic bacteria are biased towards high GC content (64%, 49%, respectively), while microaerophilic bacteria are biased towards low GC content (63%) (Fig. 4);
2. genome size:
    a. bacteria are uniformly distributed among the low/high length classes; archaea are mostly of low size
    b. in Bacteria superkingdom, there is a direct correlation between genome size and GC content;
    c. host-associated and specialized, anaerobic and microaerophilic, thermophilic and hyperthermophilic bacteria, have mostly (>70%) low genome size; aquatic, multiple, terrestrial, aerobic bacteria have mostly high genome size;
3. habitat:
    a. most of the bacteria are host-associated or multiple; most of the archaea are aquatic or specialized;
    b. terrestrial bacteria are predominantly aerobic (80%)
    c. specialized habitat bacteria are mostly thermophilic (50%); all other habitats bacteria are mostly mesophilic;
4. oxygen requirement:
    a. most of bacteria are facultative and aerobic; archaea are predominantly anaerobic;
    b. aerobic and microaerophilic bacteria are predominantly host-associated, while anaerobic bacteria are predominantly multiple;
5. optimal temperature growth:
    a. 92% of all the bacteria are mesophilic; most of the archaea are hyperthermophilic or mesophilic;
    b. higher optimal temperature (thermophilic or hyperthermophilic) is associated to specialized habitat of both bacteria and archaea and anaerobic bacteria;



(a)



(b)

Fig. 4. Habitat vs. GC content (a); Oxygen vs. GC content (b); terrestrial and aerobic bacteria are biased towards high GC content, while microaerophilic bacteria are biased towards low GC content

5.2 Association rules mined from the original NCBI data

As far as it concerns Bacteria, most of the genomes are mesophilic in temperature (more than 92%), so almost all the rules involve this element in the rule body or rule head. Some of the most reliable rules associate mesophilic bacteria with host-associated habitat and facultative anaerobic oxygen requirement (Fig. 5).

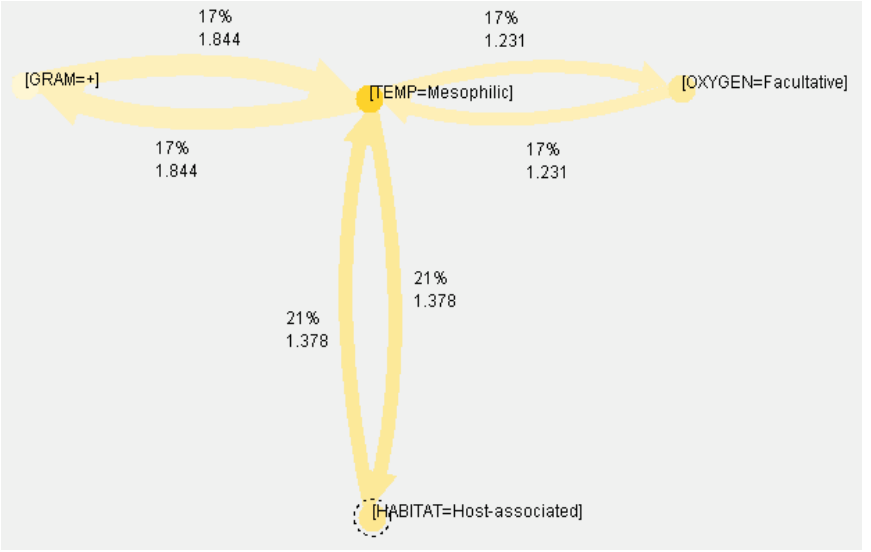
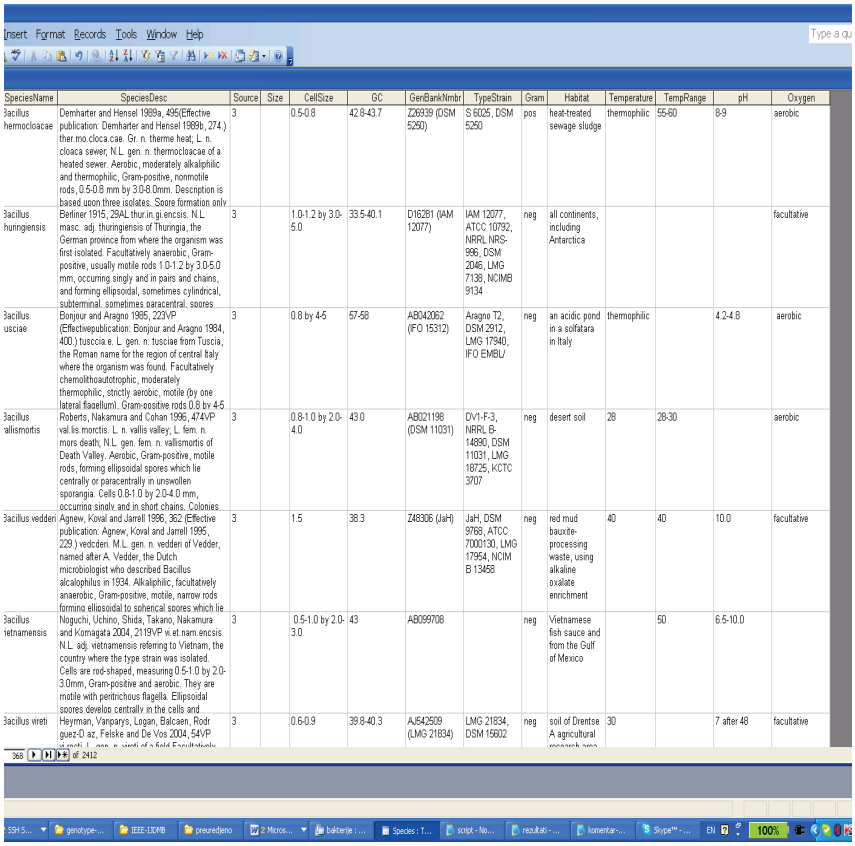


Fig. 5. Some of the most reliable association rules for the Organism_info data

5.3 Database with data extracted from text

A number of characteristics have been extracted from species description and they are stored in the table Species. But some species inherit some characteristics from the Genus they belong to, so Genus characteristics are of importance, too. The database obtained by information extraction from the Encyclopedia text contains 2412 records in the table Species and 873 records in the table Genus. An excerpt from the table Species is presented in Fig. 6. Table Genus has similar form.



SpeciesName	SpeciesDesc	Source	Size	CellSize	GC	GenBankNmbr	TypeStrain	Gram	Habitat	Temperature	TempRange	pH	Oxygen
Bacillus hermocloacae	Demharter and Hensel 1989a, 495(Effective publication: Demharter and Hensel 1989b, 274.) ther.mo.cloca.cae. Gr. n. therme heat; L. n. cloaca sewer; N.L. gen. n. thermocloacae of a heated sewer. Aerobic, moderately alkaliphilic and thermophilic, Gram-positive, nonmotile rods, 0.5-0.8 mm by 3.0-8.0mm. Description is based upon three isolates. Spore formation only	3		0.5-0.8	42.8-43.7	Z26939 (DSM 5250)	S 6025, DSM 5250	pos	heat-treated sewage sludge	thermophilic	55-60	8-9	aerobic
Bacillus thuringiensis	Berliner 1915, 29AL thur.in.gi.encsis. N.L. masc. adj. thuringiensis of Thuringia, the German province from where the organism was first isolated. Facultatively anaerobic, Gram-positive, usually motile rods 1.0-1.2 by 3.0-5.0 mm, occurring singly and in pairs and chains, and forming ellipsoidal, sometimes cylindrical, subterminal, sometimes paracentral, spores	3		1.0-1.2 by 3.0-5.0	33.5-40.1	D16281 (IAM 12077)	IAM 12077, ATCC 10792, NRRL NRS-996, DSM 2046, LMG 7138, NCIMB 9134	neg	all continents, including Antarctica				facultative
Bacillus tusciae	Bonjour and Aragno 1985, 223VP (Effectivepublication: Bonjour and Aragno 1984, 400.) tusscia.e. L. gen. n. tusciae from Tuscia, the Roman name for the region of central Italy where the organism was found. Facultatively chemolithoautotrophic, moderately thermophilic, strictly aerobic, motile (by one lateral flagellum). Gram-positive rods 0.8 by 4-5	3		0.8 by 4-5	57-58	AB042062 (IFO 15312)	Aragno T2, DSM 2912, LMG 17940, IFO EMBL/	neg	an acidic pond in a solfatara in Italy	thermophilic		4.2-4.8	aerobic
Bacillus vallismortis	Roberts, Nakamura and Cohan 1996, 474VP val.lis.mor.tis. L. n. vallis valley; L. fem. n. mors death; N.L. gen. fem. n. vallismortis of Death Valley. Aerobic, Gram-positive, motile rods, forming ellipsoidal spores which lie centrally or paracentrally in unswollen sporangia. Cells 0.8-1.0 by 2.0-4.0 mm, occurring singly and in short chains. Colonies	3		0.8-1.0 by 2.0-4.0	43.0	AB021198 (DSM 11031)	DV1-F-3, NRRL B-14890, DSM 11031, LMG 18725, KCTC 3707	neg	desert soil	28	28-30		aerobic
Bacillus vedderi	Agnew, Koval and Jarrell 1996, 362 (Effective publication: Agnew, Koval and Jarrell 1995, 229.) vedd.eri. M.L. gen. n. vedderi of Vedder, named after A. Vedder, the Dutch microbiologist who described Bacillus alcalophilus in 1934. Alkaliphilic, facultatively anaerobic, Gram-positive, motile, narrow rods forming ellipsoidal to spherical spores which lie	3		1.5	38.3	Z46306 (JaH)	JaH, DSM 9768, ATCC 700130, LMG 17954, NCIM B 13458	neg	red mud bauxite-processing waste, using alkaline oxalate enrichment	40	40	10.0	facultative
Bacillus vietnamensis	Noguchi, Uchino, Shida, Takano, Nakamura and Komagata 2004, 2119VP vi.et.nam.encsis. N.L. adj. vietnamensis referring to Vietnam, the country where the type strain was isolated. Cells are rod-shaped, measuring 0.5-1.0 by 2.0-3.0mm, Gram-positive and aerobic. They are motile with peritrichous flagella. Ellipsoidal spores develop centrally in the cells and	3		0.5-1.0 by 2.0-3.0	43	AB099708		neg	Vietnamese fish sauce and from the Gulf of Mexico		50	6.5-10.0	
Bacillus vireti	Heyrman, Vanparys, Logan, Balcaen, Rodriguez-D az, Felske and De Vos 2004, 54VP vi.reti. L. gen. n. vireti of a Solfatara. Facultatively	3		0.6-0.9	39.8-40.3	AJ542509 (LMG 21834)	LMG 21834, DSM 15602	neg	soil of Drentse A agricultural	30		7 after 48	facultative

Fig. 6. An excerpt from the table Species containing the extracted data

In case a characteristics has not been mentioned in the species description, as well as in the case the FST failed to extract the characteristic value from the species description for any reason, the corresponding genus description (if extracted) was substituted for the species characteristics in the table Species. This way, density of the table Species was significantly increased for some of the attributes – oxygen for more than 300%, temperature for about 100% , gram stain for more than 200%.

There was a need for some amount of extracted data post-processing in the form of biocuration or manual data unification and harmonization. In general, description of a species habitat had to be mapped into the set of habitat modalities from the NCBI database: {host-associated, specialized, multiple, aquatic, terrestrial}; for example, from the text "desert soil", extracted by the FST for the attribute *Habitat* for the species *Bacillus Vallismortis*, the value "terrestrial" had to be inferred manually. In some other cases, literature had to be consulted, such as "heat-treated sewage-sludge" for the species *Bacillus Thermocloaceae*.

The number of extracted characteristics for Species (Genus – respectively), are as follows: 410 (554) for Oxygen, 485 (738) for Gram stain, 711 (190) for pH, 1616 (284) for Habitat, 455 (257) for Temperature, 638 (170) for TempRange. Information on a number of species from the table Species were already present in the NCBI Organism_info table, but for quite a lot they were not. Most of the values extracted coincide with those in the NCBI database (if present for the corresponding species), but in some cases they differ. Fig. 7. represents the total number of species extracted along with the portion of them present and absent from the NCBI Organism_info table, as well as the number of values for different characteristics (oxygen requirement, habitat, temperature range, gram stain) extracted from the text that are present in the Organism_info and coincide with them or differ from them.

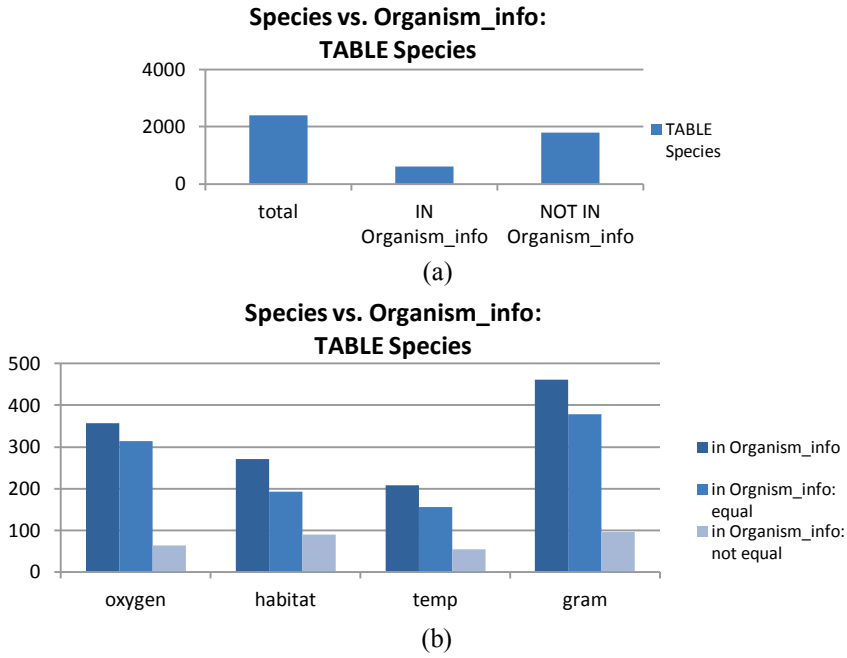


Fig. 7. Number of species extracted from the encyclopedia text (table Species) (a); number of specific values for different characteristics extracted from the encyclopedia text (table Species), present in the Organism_info table (NCBI), as well as the number of coinciding and differing values

5.4 Integration of extracted data with NCBI data

Data extracted from the encyclopedia are integrated with NCBI data in two ways:

a) Filling empty cells (missing characteristics values) in the Organism_info (NCBI) table by extracted data (table Species extended by Genus data); total number of records in the Organism_info table remains unchanged (table Organism_info_int). Fig. 8. represents the number of values defined for different characteristic in the original NCBI Organism_info table and in it enhanced version Orginsm_info_int.

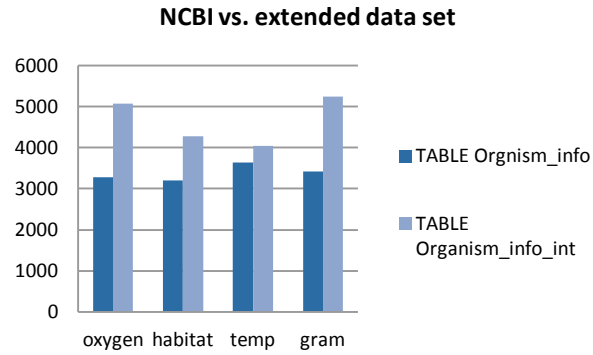


Fig. 8. Number of values defined for different characteristic in the original NCBI Organism_info table and in its extended version Orginsm_info_int

b) Enlarging the number of records (species) in the Organism_info table by adding data on new species extracted from the encyclopedia text; this enlargement implies elimination of those characteristics from the Organism_info table that were not extracted from the encyclopedia text (or not even present there) - table Species_int. The new table Species_int thus contains common characteristics of the two tables (Species and Organism_info_int) only, and it is obtained by projection of the tables union to common attributes. Fig. 9. represents the overall effect of the two types of data integration - increasing number of defined values for the selected attributes (characteristics) in the original NCBI Organism_info table, enhanced Organism_info by adding characteristics extracted from the encyclopedia text (Organism_info_int) and in the table obtained by adding data on newly extracted species from the encyclopedia text (table Species_int).

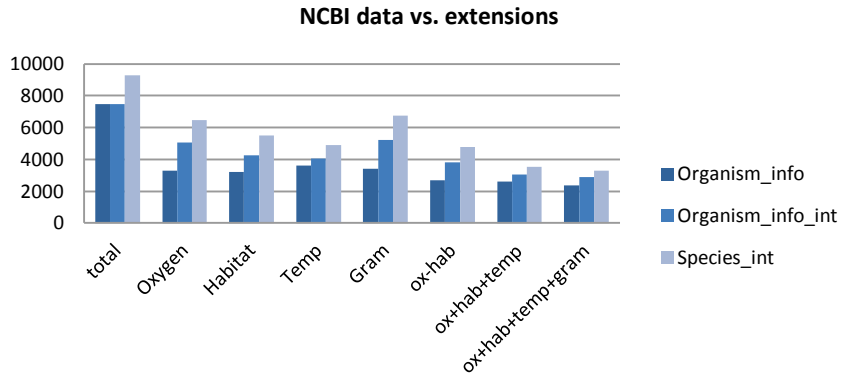


Fig. 9. Number of total records and values for different chrcteristics in the original NCBI data and its to extensions obtained by extracting data from text

5.5 Association rule mining for the integrated data

Mining association rules involving phenotypic data only (habitat, oxygen, temperature, gram stain), with enriched set of organisms and attribute values - extracted species form text (table *Species_int*), discard some of the less reliable and multiply related characteristics and add new associations (some of them presented in Fig. 10.).

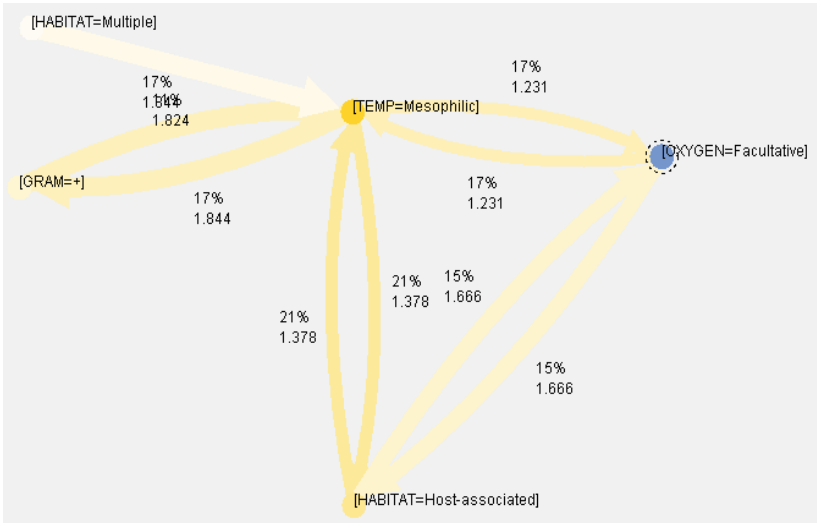


Fig. 10. New association rules mined from extended data collection: Host-associated-Facultative anaerobic prokaryotes

6 Conclusion

By extracting data from unstructured or semi-structured documents, a significant enlargement of existing databases can be obtained. In this paper, it has been demonstrated by integrating the database of prokaryote characteristics with the extracted data from the encyclopedia of bacteriology. But the same approach is applicable to many other specific areas and tasks.

Intended enrichment in association rules mined over the enriched data collection was only partly successful. It is mainly due to sparse structured (NCBI) data, and still rather sparse integrated collection. The results suggest that integration with other microbial databases which would provide for more populated data collection, is necessary. It would provide for even richer set of characteristics, such as genotypic, phenotypic, structural, protein function and structure, RNA secondary structure, etc.

Finally, the results obtained suggest that association rules may not be the most appropriate method of analyzing relationships and dependencies among the organism characteristics, so that multivariate analysis may be the next step and even the method of choice.

Acknowledgment

The work presented has been supported by the Ministry of Education and Science, Republic of Serbia, Projects No. 174021.

References

1. http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html
2. Jim K, Parmar K, Singh M, Tavazoie S. A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes, Genome Research, 14, 2004, 109-115
3. Burra PV, Kalmar L, Tompa P. Reduction in Structural Disorder and Functional Complexity in the Thermal Adaptation of Prokaryotes. PLoS ONE 5(8): e12069. doi:10.1371/journal.pone.0012069 (2010)
4. Goh CS, Gianoulis TA, Liu Y, Li J, Paccanaro A, Lussier YA, Gerstein M. Integration of curated databases to identify genotype-phenotype associations, *BMC Genomics,* 7, 2006, pp.257–257.,
5. Guo FB, Lin H, Huang J. A plot of G + C content against sequence length of 640 bacterial chromosomes shows the points are widely scattered in the upper triangular area, Chromosome Research (2009) 17:359–364
6. Wu H, Zhang Z, Huand S, Yu J. On the molecular mechanism of GC content variation among eubacterial genomes, Biology Direct2012,7:2
7. Rocha EPC, Danchin . Base composition bias might result from competition for metabolic resources TRENDS in Genetics,18(6), 291-294, (2002)
8. Sabath N, Ferrada E, Barve A, Wagner A. Growth Temperature and Genome Size in Bacteria Are Negatively Correlated, Suggesting Genomic Streamlining During Thermal Adaptation, Genome Biol. Evol.5(5), 966–977
9. Bohlin J, Snipen L, Hardy SP, Kristoffersen AB, Lagesen K, Dønsvik T, Skjerve E, Ussery DW. Analysis of intra-genomic GC content homogeneity within prokaryotes BMC Genomics, 11, 464, (2010)
10. Romero H, Pereira E, Naya H, Musto H. Oxygen and Guanine–Cytosine Profiles in Marine Environments J. Mol. Evol., 69, 203–206, (2009)
11. Lin H, HuangY, Zhang S. Correlation Between Genome Size and GC Content in Prokaryotes with Different Lifestyles, Acta Scientiarum Naturalium Universitatis Sunyatseni, ISSN: 0529-6579 , 2011, 50(3): 90-93.
12. Hurstl LD, Merchant AR. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes, Proc Biol Sci. Mar 7, 2001; 268(1466): 493–497. doi: 10.1098/rspb.2000.1397
13. Musto H, Naya H, Zavala Z, Romero H, Alvarez-Valin F, Bernardi G. Genomic GC level, optimal growth temperature, and genome size in prokaryotes, Biochem Biophys Res Commun. 2006;347(1):1-3
14. Mann S, ChenYPP. Bacterial genomic G+C composition-eliciting environmental adaptation, *Genomics*, Vol. 95, 2010, pp.7{15.
15. Creighton C, Hanash S. Mining gene expression databases for association rules, *Bioinformatics*, Vol. 19, 2003, pp.79{86.
16. Georgii E, Richter L, Ruckert, U. and Kramer, S. Analyzing microarray data using quantitative association rules, *Bioinformatics*, Vol. 21(Suppl 2), 2005:II123-II129.
17. Carmona-Saez P, Chagoyen M, Rodriguez A, Trelles O, Carazo JM, Pascual-Montano A. Integrated analysis of gene expression by association rules discovery, *BMC Bioinformatics*, Vol. 7, (2006) doi:10.1186/1471-2105-7-54.
18. Martinez R, Pasquier C, Pasquier N. GenMiner: Mining Informative Association Rules from Genomic Data, *IEEE International Conference 2007 on Bioinformatics and Biomed-*

icine (IEEE BIBM 2007), pp 15-22.
19. Martinez R, Pasquier N, Pasquier C. Mining Association Rule Bases from Integrated Genomic Data and Annotations, *In Masulli, F. et al. (Eds.) CIBB 2008. LNBI 5488 Springer-Verlag Berlin Heidelberg*, pp 78-90.
20. Gyenesei A, Wagner U, Barkow-Oesterreicher S, Stolte E, Schlapbach R. Mining co-regulated gene profies for the detection of functional associations in gene expression data, *Bioinformatics*, vol. 23, (2007) pp.1927-1935.
21. Tamura M, D'haeseleer P. Microbial genotype-phenotype mapping by class association rule mining. Bioinformatics. 1;24(13):1523-9. (2008)
22. MacDonald NJ, Beiko RG. Efficient learning of microbial genotype-phenotype association rules, Bioinformatics, 26, 2010, pp. 1834-1840.
23. Korbel J, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P. Systematic association of genes to phenotypes by genome and literature mining, PLoS Biol, 3, 2005, pp. 134-134.
24. Jimeno-Yepes A, Berlanga-Llavori R, Rebholz-Schuhmann D. Exploitation of ontological resources for scientific literature analysis: Searching genes and related diseases,*Engineering in Medicine and Biology Society, EMBC 2009, Annual International Conference of the IEEE*,pp. 7073–7078.
25. Pajić VS, Pavloviç-Lazetić GM, Beljanski MV, Brandt BW, Pajić MB. Towards a database for genotype-phenotype association research: mining data from encyclopaedia., Int J Data Min Bioinform. 2013;7(2):196-213.
26. Pajic V, Pavlovic Lazetic G, Pajic M. Information Extraction from Semi-structured Resources: A Two-Phase Finite State Transducers Approach, Lecture Notes in Computer Science, 6807, ISSN 0302-9743, Springer Berlin Heidelberg (2011)
27. Pavlovic-Lazetic G, Pajic V, Mitic N, Kovacevic J, Beljanski M. Mining Associations for Organism Characteristics in Prokaryotes – an Integrative Approach, International work-conference on bioinformatics and biomedical engineering, IWBBIO 2014, April 7-9, Granada, Spain
28. http://cmr.jcvi.org/tigr-scripts/CMR/CmrHomePage.cgi
29. http://www.cbs.dtu.dk/services/GenomeAtlas/
30. http://img.jgi.doe.gov/cgi-bin/w/main.cgi
31. http://www.patricbrc.org
32. Garrity G. ed. (2005) *Bergey's Manual of Systematic Bacteriology, Volume 2: The Proteobacteria,*ISBN 978-0-387-95040-2.
33. Whitman W.B, ed (2009) *Bergey's Manual of Systematic Bacteriology, Volume 3: The Firmicutes,* ISBN: 978-0-387-95041-9
34. Whitman W.B, ed (2010) *Bergey's Manual of Systematic Bacteriology, Volume 4: The Bacteroidetes, Spirochaetes, Tenericutes (Mollicutes), Acidobacteria, Fibrobacteres, Fusobacteria, Dictyoglomi, Gemmatimonadetes, Lentisphaerae, Verrucomicrobia, Chlamydiae, and Planctomycetes*, ISBN: 978-0-387-95042-6
35. Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar HE, Moran NA, Hattori M. The 160-kilobase genome of the bacterial endosymbiont Carsonella, *Science*, Vol. 314, 2006, p.267.
36. Koonin EV, Wolf YI. Genomics of Bacteria and Achaea: the emerging dynamic view of the prokaryotic world, *Nucleic Acids Research*, Vol. 36, 2008, pp.6688-6719.
37. Gardiner K. Base composition (GC composition, GC richness), *In Dictionary of Bioinformatics and Computational Biology (eds. Hancock, J.M.and Zvelebil, M.J.), Wiley-Liss, Hoboken, New Jersey*, 2004, p.40.

38. Mann S, ChenYPP. Bacterial genomic G+C composition-eliciting environmental adaptation, *Genomics*, Vol. 95, 2010, pp.7-15.
39. Schlegel H.G, Jannasch HW. Prokaryotes and Their Habitats, *in The Prokaryotes, Martin Dworkin (Ed.),Springer Science{Business Media, LLC,New York*, Vol. 1, 2006, pp.137-184.
40. Paumier, S. Unitex 1.2 User Manual, Université de Marne-la-Vallée, (2006). http://www-igm.univ-mlv.fr/'unitex/UnitexManual.pdf

Matrix Genetics: Algebra of Projection Operators, Cyclic Groups and Inherited Ensembles of Biological Cycles

Sergey Petoukhov [a]
Laboratory of biomechanical systems of the Mechanical Engineering Research Institute of the Russian Academy of Sciences, Moscow, Russia

Abstract

This work is devoted to applications of projection operators to study molecular-genetic systems and some inherited physiological phenomena. The author analyzes matrix representations of ensembles of molecular-genetic elements in a form of a tensor family of matrices [C U; A G]$^{(n)}$ where C means cytosine, U - uracil, A - adenine, G - guanine, (n) - tensor power. The (8*8)-matrix [C U; A G]$^{(3)}$ contains all 64 triplets in a strong order. The nature has divided the set of 64 triplets into two equal subsets: 32 triplets have "strong roots" and 32 triplets have "weak roots". A disposition of triplets with strong roots and weak roots in the

[a] e-mail address: spetoukhov@gmail.com

matrix [C U; A G]$^{(3)}$ gives such symmetrical mosaic that a mosaic of each column of this matrix has meander-like character and coincides with one of Rademacher functions. By replacing each of triplets with a strong (weak) root with number "+1" ("-1"), we receive a numeric matrix R as a "Rademacher representation" of the symbolic matrix [C U; A G]$^{(3)}$. This matrix R is a sum of such 8 sparse matrices that each sparse matrix is an oblique projection operator (it satisfies the criterion $P^2 = P$). Combinations of these "genetic" (8*8)-projectors in a form of sums of two or more projectors lead to sets of cyclic groups and other interesting mathematical objects, which are used by the author to model inherited ensembles of biological phenomena including cyclic processes. Some relations of hypercomplex numbers with the Rademacher (8*8)-matrix R and its algorithmic extensions into $(2^N * 2^N)$-matrices are revealed. Projectors are used widely in mathematics, physics (including quantum mechanics), chemistry, informatics, logics, etc. Our results give evidences that the notion of "projection operators"can be one of useful notions and instruments in the field of bioinformatics and mathematical biology. We believe that living matter is an algebraic essence in its informational fundamentals. [b]

[b]Work supported by the Russian State contract No. P377 from 30.07.2009

It is known that our visual perception is based on projections of external objects on the retina of our eyes. But any organism is a single entity, and our work shows that the importance of the projection operators for living organisms is not confined to this separate fact and that these operators play a fundamental role to bioinformatics at all. In mathematics, such operations of projections are expressed by means of square matrices, which are called "projection operators" or "projectors" (http://en.wikipedia.org/wiki/Projection_(linear_algebra)). The necessary and sufficient condition that a matrix P is a projection operator is given by the criterion: $P^2 = P$. Two main types of projectors exist: orthogonal projectors, which are expressed by means of symmetric matrices, and oblique projectors, which are expressed by means of non-symmetric matrices. Oblique projectors are main objects in our study because of their relations with the molecular-genetic system.

Science has led to a new understanding of life itself: "*Life is a partnership between genes and mathematics*" [1]. But what kind of mathematics can be a partner for the genetic coding system? This article shows some evidences that algebra of projectors can be one of main parts of such mathematics.

In accordance with Mendel's laws of independent inheritance of traits, information from the micro-world of genetic molecules dictates constructions in the macro-world of living organisms under strong noise and interference. This dictation is realized by means of unknown algorithms of multi-channel noise-immunity coding. For example, in human organism, his skin color, eye color and hair color are inherited genetically independently of each other. It is possible if appropriate kinds of information are conducted via independent informational channels and if a multi-parametric system of living organism contains sub-systems with a possibility of a selective control or a selective coding of processes in them. So, any living organism is a multi-parametric algorithmic machine of multi-channel noise-immunity coding with ability to a selective control and coding of different sub-systems of this multi-parametric machine (a model approach to multi-parametric systems with a selective control of their sub-systems is presented in this article). This machine works in conditions of ontogenetic development of the organism when its multi-parametric organization is complicated step by step.

To understand such genetic machine, it is appropriate to use the theory of noise-immunity coding and transmission of digital information, taking into account the discrete nature of the genetic code. In this theory, mathematical matrices have the basic importance. The use of matrix representations and analysis in the study of phenomenological features of molecular-genetic ensembles has led to the development of a scientific direction under a name "Matrix Genetics" [2-4].

Matrices of genetic duplets and triplets. The 4-letter alphabet of RNA (adenine A, cytosine C, guanine G and uracil U) can be represented in a form of the symbolic (2*2)-matrix [C U; A G] as a kernel of the tensor (or Kronecker) family of symbolic matrices $[C\ U;\ A\ G]^{(n)}$, where (n) means a tensor power (Figure 1). Inside this family, this 4-letter alphabet of monoplets is connected with the alphabet of 16 duplets and 64 triplets by means of the second and third tensor powers of the kernel matrix: $[C\ U;\ A\ G]^{(2)}$ and $[C\ U;\ A\ G]^{(3)}$, where all duplets and triplets are disposed in a strict order (Figure 1). We begin with the alphabet A, C, G, U of RNA here because mRNA-sequences of triplets define protein sequences of amino acids in a course of their reading in ribosomes.

The matrix $[C\ U;\ A\ G]^{(3)}$ on Figure 1 contains not only 64 triplets but also amino acids and stop-codons encoded by the triplets in the case of the Vertebrate mitochondrial genetic code that is the most symmetrical among known variants of the genetic code (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi). Let us explain the black-and-white mosaics of $[C\ U;\ A\ G]^{(2)}$ and $[C\ U;\ A\ G]^{(3)}$ (Figure 1) which reflect important features of the genetic code. These features are connected with a specificity of reading of mRNA-sequences in ribosomes to define protein sequences of amino acids.

C	U
A	G

CC	CU	UC	UU
CA	CG	UA	UG
AC	AU	GC	GU
AA	AG	GA	GG

CCC Pro	CCU Pro	CUC Leu	CUU Leu	UCC Ser	UCU Ser	UUC Phe	UUU Phe
CCA Pro	CCG Pro	CUA Leu	CUG Leu	UCA Ser	UCG Ser	UUA Leu	UUG Leu
CAC His	CAU His	CGC Arg	CGU Arg	UAC Tyr	UAU Tyr	UGC Cys	UGU Cys
CAA Gln	CAG Gln	CGA Arg	CGG Arg	UAA Stop	UAG Stop	UGA Trp	UGG trp
ACC Thr	ACU Thr	AUC Ile	AUU Ile	GCC Ala	GCU Ala	GUC Val	GUU Val
ACA Thr	ACG Thr	AUA Met	AUG Met	GCA Ala	GCG Ala	GUA Val	GUG Val
AAC Asn	AAU Asn	AGC Ser	AGU Ser	GAC Asp	GAU Asp	GGC Gly	GGU Gly
AAA Lys	AAG Lys	AGA Stop	AGG Stop	GAA Glu	GAG Glu	GGA Gly	GGG Gly

Figure 1. The first three representatives of the tensor family of RNA-alphabetic matrices $[C\ U;\ A\ G]^{(n)}$. Black color marks 8 strong duplets in the matrix $[C\ U;\ A\ G]^{(2)}$ (at the top) and 32 triplets with strong roots in the matrix $[C\ U;\ A\ G]^{(3)}$ (bottom). 20 amino acids and stop-codons, which correspond to the triplets, are also shown in the matrix $[C\ U;\ A\ G]^{(3)}$ for the case of the Vertebrate mitochondrial genetic code

A combination of letters on the two first positions of each triplet is usually termed as a "root" of this triplet [5, 6]. Modern science recognizes many variants (or dialects) of the genetic code, data about which are shown on the NCBI's website http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi. 19 variants (or dialects) of the genetic code exist that differ one from another by some details of correspondences between triplets and objects encoded by them (these dialects are known at July 10, 2013, but perhaps later their list will be increased). Most of these dialects (including the so called Standard Code and the Vertebrate Mitochondrial Code) have the symmetric general scheme of these correspondences, where 32 "black" triplets with "strong roots" and 32 "white" triplets with "weak" roots exist. In this basic scheme, the set of 64

triplets contains 16 subfamilies of triplets, every one of which contains 4 triplets with the same two letters on the first positions (an example of such subsets is four triplets CAC, CAA, CAT, CAG with the same two letters CA on their first positions). In the described basic scheme, the set of these 16 subfamilies is divided into two equal subsets. The first subset contains 8 subfamilies of triplets, a coding value of which is independent on a letter on their third position: (CCC, CCT, CCA, CCG), (CTC, CTT, CTA, CTG), (CGC, CGT, CGA, CGG), (TCC, TCT, TCA, TCG), (ACC, ACT, ACA, ACG), (GCC, GCT, GCA, GCG), (GTC, GTT, GTA, GTG), (GGC, GGT, GGA, GGG). An example of such subfamilies is the four triplets CGC, CGA, CGT, CGC, all of which encode the same amino acid Arg. The 32 triplets of the first subset are termed as "triplets with strong roots" [5, 6]. The following duplets are appropriate 8 strong roots for them: CC, CT, CG, AC, TC, GC, GT, GG („strong duplets"). Each of these 32 triplets and 8 strong duplets are marked by black color in the matrices $[C\ U;\ A\ G]^{(3)}$ and $[C\ U;\ A\ G]^{(2)}$ on Figure 1.

The second subset contains 8 subfamilies of triplets, the coding value of which depends on a letter on their third position: (CAC, CAT, CAA, CAG), (TTC, TTT, TTA, TTG), (TAC, TAT, TAA, TAG), (TGC, TGT, TGA, TGG), (AAC, AAT, AAA, AAG), (ATC, ATT, ATA, ATG), (AGC, AGT, AGA, AGG), (GAA, GAT, GAA, GAG). An example of such subfamilies is four triplets CAC, CAA, CAT, CAC, two of which (CAC, CAT) encode the amino acid His and the other two (CAA, CAG) encode another amino acid Gln. The 32 triplets of the second subset are termed as "triplets with weak roots" [5, 6]. The following duplets are appropriate 8 weak roots for them: CA, AA, AT, AG, TA, TT, TG, GA („weak duplets"). Each of these 32 triplets and 8 weak duplets are marked by white color in matrices $[C\ U;\ A\ G]^{(3)}$ and $[C\ U;\ A\ G]^{(2)}$ on Figure 1.

What secrets of living matter are hidden in these symmetric black-and-white matrices? The most important fact is the following: from the point of view of its black-and-white mosaic, each of columns of genetic matrices $[C\ U;\ A\ G]^{(2)}$ and $[C\ U;\ A\ G]^{(3)}$ has a meander-like character and coincides with one of Rademacher functions that form orthogonal systems and well known in discrete signals processing. These functions contain elements "+1" and "-1" only. Due to this fact, one can construct Rademacher representations of the symbolic genomatrices $[C\ U;\ A\ G]^{(2)}$ and $[C\ U;\ A\ G]^{(3)}$ (Figure 1) by means of the following operation: each of black duplets and of black triplets is replaced by number "+1" and each of white duplets and white triplets is replaced by number "-1". This operation leads immediately to the matrices R_4 and R_8 (Figure 2), that are the Rademacher representations of the phenomenological genetic matrices $[C\ U;\ A\ G]^{(2)}$ and $[C\ U;\ A\ G]^{(3)}$.

$$\mathbf{R_4} = \begin{vmatrix} 1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{vmatrix} ; \qquad \mathbf{R_8} = \begin{vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & 1 & 1 \end{vmatrix}$$

Figure 2. Numeric matrices R_4 and R_8 which are related with phenomenology of the genetic coding system

Every of these matrices R_4 and R_8 can be decomposed into sum of sparse matrices, each of which contains only one non-zero column. Such decomposition can be conditionally termed a «column decomposition». Figure 3 shows an example of such decomposition for the matrix R_4.

$$R_4 = c_0+c_1+c_2+c_3 = \begin{vmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{vmatrix} + \begin{vmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{vmatrix} + \begin{vmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \end{vmatrix} + \begin{vmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{vmatrix}$$

Figure 3. The «column decomposition» of the Rademacher matrix R_4

Each of these sparse matrices c_0, c_1, c_2 and c_3 is a projection operator because it satisfies the criterion of projectors $P^2=P$ (for example, $c_0^2=c_0$, etc). Every of these projectors is an oblique (non-orthogonal) projector because it is expressed by means of a non-symmetrical matrix. By analogy, each of 8 sparse matrices, which arises in the similar "column decomposition" of the Rademacher (8*8)-matrix R_8, is also oblique projector (see details in [3]).

Let us examine sums of pairs of the projectors (c_0+c_1) and (c_2+c_3). Figure 4 shows that each of these sums can be decomposed in an appropriate way into a set of two matrices e_0 and e_1 (e_2 and e_3 correspondingly), which is closed relative to multiplication and which defines a known multiplication table of complex numbers. It means that these (4*4)-matrices (c_0+c_1) and (c_2+c_3) represent complex numbers with unit coordinates in different planes (x_0,x_1) and (x_2,x_3) inside a 4-dimensional space (x_0,x_1,x_2,x_3).

$$c_0+c_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ -1 & -1 & 0 & 0 \end{bmatrix} = e_0+e_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}$$

		e_0	e_1
;	e_0	e_0	e_1
	e_1	e_1	$-e_0$

$$c_2+c_3 = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & 1 \end{bmatrix} = e_2+e_3 = \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}$$

		e_2	e_3
;	e_2	e_2	e_3
	e_3	e_3	$-e_2$

Figure 4. Decompositions of (4*4)-matrices (c_0+c_1) and (c_2+c_3) into sum of two matrices (e_0+e_1) and (e_2+e_3). The table shows direct relations of these matrices with matrix representations of 2-parametric complex numbers with unit coordinates. The right column of this Figure shows multiplication tables of complex numbers for appropriate sets of matrices (e_0, e_1) and (e_2, e_3)

A similar situation holds true for a column decomposition of the Rademacher (8*8)-matrix R_8, where 4 different complex numbers with unit coordinates are hidden inside an appropriate 8-dimensional space [3]. Moreover, (2^N*2^N)-matrices, which contain 2^{N-1} 2-dimensional planes of complex numbers inside 2^N-dimensional spaces, can be constructed by means of the algorithm $R_4 \otimes [1\ 1;\ 1\ 1]^{(n)}$ where $\otimes$ means tensor multiplication and (n) means tensor exponentiation of the matrix representation [1 1; 1 1] of hyperbolic number with unit coordinates [3]. But this brief paper can explain a general situation with genetic matrices and their projectors only by means of an example of the genetic (4*4)-matrix R_4.

Using the basic elements of complex numbers e_0, e_1 and e_2, e_3 from Figure 4, one can create the following two subsets of matrices: $M_L=a_0*e_0+a_1*e_1$ and $M_R=a_2*e_2+a_3*e_3$ (Figure 5). Each of these subsets is a matrix representation of 2-parametric complex numbers in an appropriate 2-dimensional plane inside a 4-dimensional space. Each of these subsets M_L or M_R doesn't contain usual unit matrix [1 0 0 0; 0 1 0 0; 0 0 1 0; 0 0 0 1] but contains matrices e_0 or e_2, each of which plays there a role of unity matrix. Figure 5 shows also appropriate expressions of inverse complex numbers M_L^{-1} or M_R^{-1}.

$$M_L=a_0*e_0+a_1*e_1=\begin{matrix} a_0 & 0 & 0 & 0 \\ 0 & a_0 & 0 & 0 \\ 0 & -a_0 & 0 & 0 \\ -a_0 & 0 & 0 & 0 \end{matrix} + \begin{matrix} 0 & a_1 & 0 & 0 \\ -a_1 & 0 & 0 & 0 \\ a_1 & 0 & 0 & 0 \\ 0 & -a_1 & 0 & 0 \end{matrix} = \begin{matrix} a_0 & a_1 & 0 & 0 \\ -a_1 & a_0 & 0 & 0 \\ a_1 & -a_0 & 0 & 0 \\ -a_0 & -a_1 & 0 & 0 \end{matrix}$$

$$M_R=a_2*e_2+a_3*e_3=\begin{matrix} 0 & 0 & 0 & -a_2 \\ 0 & 0 & -a_2 & 0 \\ 0 & 0 & a_2 & 0 \\ 0 & 0 & 0 & a_2 \end{matrix} + \begin{matrix} 0 & 0 & a_3 & 0 \\ 0 & 0 & 0 & -a_3 \\ 0 & 0 & 0 & a_3 \\ 0 & 0 & -a_3 & 0 \end{matrix} = \begin{matrix} 0 & 0 & a_3 & -a_2 \\ 0 & 0 & -a_2 & -a_3 \\ 0 & 0 & a_2 & a_3 \\ 0 & 0 & -a_3 & a_2 \end{matrix}$$

$$M_L^{-1} = (a_0^2+a_1^2)^{-1} * \begin{matrix} a_0 & -a_1 & 0 & 0 \\ a_1 & a_0 & 0 & 0 \\ -a_1 & -a_0 & 0 & 0 \\ -a_0 & a_1 & 0 & 0 \end{matrix} \quad ; \quad M_R^{-1} = (a_2^2+a_3^2)^{-1} * \begin{matrix} 0 & 0 & -a_3 & -a_2 \\ 0 & 0 & -a_2 & a_3 \\ 0 & 0 & a_2 & -a_3 \\ 0 & 0 & a_3 & a_2 \end{matrix}$$

Figure 5. Each of (4*4)-matrices M_L or M_R represents complex numbers in an appropriate plane of a 4-dimensional space. Lower level: expressions of inverse complex numbers M_L^{-1} and M_R^{-1}. Here a_0, a_1, a_2, a_3 are real numbers.

Product of two matrices from different subsets M_L and M_R is non-commutative. Product of any two matrices from the same subset M_L (or from M_R) generates a matrix from the same subset. This property allows a selective management (or encoding) of processes in separate planes (x_0,x_1) and (x_2,x_3) by independent way. The same property holds true for many planes of complex numbers in 2^N-dimensional spaces, which is managed by means of (2^N*2^N)-matrix operators created on the base of the mentioned algorithm $R_4\otimes[1\ 1;\ 1\ 1]^{(n)}$ [3]. If entries of such matrix operator are functions of time, the operator describes the life of a multi-parametric system in time. Such (2^N*2^N)-matrix operators can be used to simulate 2^N-parametric systems, whose subsystems are managed (or encoded) independently to each other; these subsystems can be also managed in interconnected manner in special cases when functions in matrix components are interconnected.

Any living organism can be considered as a multi-parametric system with many sub-systems, whose functional interrelations are expressed in more or less extent; this system and its sub-systems are genetically inherited and consequently should be modeled in a connection with genetic formalisms. The described (2^N*2^N)-matrix operators, which are based on genetic projectors, allow modeling many inherited ensembles of living organisms including ensembles of cyclic processes, phyllotaxis patterns, etc.

Let us show an example of the simplest case of (4*4)-matrix. The iterative action of the (4*4)-operator $2^{-0.5}*(c_0+c_1)$, where (c_0+c_1) is the sum of genetic projectors from Figure 4, on an arbitrary 4-dimensional vector $X=[x_0,x_1,x_2,x_3]$ generates new vectors $X*(2^{-0.5}*(c_0+c_1))^N$, which belong only to the plane (x_0,x_1) and have zero values on coordinate axes x_2 and x_3. Secondly, exponentiation of this operator $2^{-0.5}*(c_0+c_1)$ forms a cyclic group of transformation, whose period is equal to 8: $(2^{-0.5}*(c_0+c_1))^N = (2^{-0.5}*(c_0+c_1))^{N+8}$. The similar situation holds true for another (4*4)-matrix operator $2^{-0.5}*(c_2+c_3)$, where (c_2+c_3) is also the sum of genetic projectors from Figure 4, but in this case its action on the vector $X=[x_0,x_1,x_2,x_3]$ generates new vectors $X*(2^{-0.5}*(c_2+c_3))$, which belong only to the plane (x_2,x_3) and have zero values on coordinate axes x_0 and x_1. A similar cyclic group also arises in this case: $(2^{-0.5}*(c_2+c_3))^N = (2^{-0.5}*(c_2+c_3))^{N+8}$. These two cyclic groups $(2^{-0.5}*(c_0+c_1))^N$ and $(2^{-0.5}*(c_0+c_1))^N$ allow modeling a 4-parametric system, which consists of two 2-parametric sub-systems with a cyclic behavior in time. Using oblique projectors from the column decomposition of (2^N*2^N)-matrix operators $R_4 \otimes [1\ 1;\ 1\ 1]^{(n)}$, one can obtain as many cyclic groups as needed to model big ensembles of cyclic processes in a connection with genetic phenomenology. If components in the described matrix operators are functions in time, we have operators, which are changed in time and which describe multi-parametric systems, whose sub-systems can be changed in time independently to each other.

Any living organism is an object with a huge ensemble of inherited cyclic processes, which form a hierarchy at different levels. Even every protein is involved in a cycle of the "birth and death," because after a certain time it breaks down into its constituent amino acids and they are then collected into a new protein. According to chrono-medicine and bio-rhythmology, various diseases of the body are associated with disturbances (dis-synchronization) in these cooperative ensembles of its cycles. All inherited physiological subsystems of the body should be agreed with the structural organization of genetic coding for their coding and transmission to descendants; in other words, they bear the stamp of its features. We develop a "genetic biomechanics", which studies deep coherence between inherited physiological systems and molecular-genetic structures. In this scientific field the mentioned matrix formalisms and cyclic groups are used, in particularly, to model human and animal gaits that are genetically inherited. Really, many gaits (which are based on cyclic movements of the limbs and the corresponding muscle actuators) have genetically inherited character. For example, newborn turtles and crocodiles, when they hatched from eggs, crawl with quite coordinated movements to water without any training from anybody; centipedes crawl by means of coordinated movements of a great number of their legs on the basis of inherited algorithms of control of legs. One

should emphasize that, in the previous history, gaits and locomotion algorithms were studied in biomechanics of movements without any connection with the structures of genetic coding and with inheritance of unified control algorithms.

The exclusion principle for dialects of the genetic code. Discovering of exclusion principles of nature is a significant task of mathematical natural science (the exclusion principle by Pauli in quantum mechanics is one of examples). Now the author formulates the exclusion principle of evolution of dialects of the genetic code, which has been discovered in study of described genetic projectors.

The site http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi contains all known 19 dialects of the genetic code presented at July 10, 2013. The author has analyzed black-and-white genetic matrices $[C\ U;\ A\ G]^{(3)}$, where black cells correspond to triplets with strong roots and white cells correspond to triplets with weak roots, for each of these 19 dialects. The results are the following. The vast majority of dialects (13 dialect from the set of 19 dialects) possess the same black-and-white mosaics as on Figure 1 though some triplets have different code meanings in different dialects. Only 6 dialects have their matrix $[C\ U;\ A\ G]^{(3)}$ with atypical black-and-white mosaics: Invertebrate Mitochondrial Code; Echinoderm and Flatworm Mitochondrial Code; Alternative Yeast Nuclear Code; Alternative Flatworm Mitochondrial Code; Trematode Mitochondrial Code; Scenedesmus Obliquus Mitochondrial Code.

By analogy with the Rademacher presentation R_8 (Figures 2), one can again replace black triplets by elements «+1» and white triplets by elements «-1» to receive numeric representations of these genetic matrices $[C\ U;\ A\ G]^{(3)}$ for all dialects. Such numeric representations of genetic matrices can be conditionally called as «±1-representations». The phenomenologic fact has been revealed: this numeric «±1-representation» of matrices $[C\ U;\ A\ G]^{(3)}$ for every of 19 dialects is decomposed into a sum of 8 sparse (8*8)-matrices, each of which is a column projector. In other words, algebra of projectors shows an existence of an algebraic invariant of biological evolution. So the phenomenologic exclusion principle can be formulated for evolutionary changes of dialects of the genetic code: it is forbidden for biological evolution to violate a separation of the set of 64 triplets into two subsets of triplets with strong and weak roots (black and white triplets) in such way that - for a new dialect - a black-and-white mosaic of the genetic matrix $[C\ U;\ A\ G]^{(3)}$ in its «±1-representation» ceases to be a sum of 8 column projectors [3].

The generalization of hypercomplex numbers. Let us consider sum of matrices M_L and M_R from Figure 5. Figure 6 shows this sum M=ML+MR. One can see a symmetrical character of this matrix M: its upper half is mirror-

antisymmetric to its lower half (it resembles the mirror-antisymmetric character of the matrix [C U; A G][(2)] on Figure 1); more precisely, any two components of the matrix, which are located mirror-symmetrical in its upper and lower halves, are the same, but have the opposite sign. Matrices of this type M can be added, subtracted and multiply with each other, receiving in the result a matrix of the same type (division operation in this set of matrices isn't determined). This means that the set of such matrices has important properties of a system of 4-dimensional numbers.

Figure 6 shows a decomposition of the matrix M into a form M=E*K+P*K, where E is unit matrix; P is the permutation matrix with minus signs of its non-zero entries and with the property P^2=E; K is the (4*4)-matrix, the upper half of which is taken from the matrix M and the lower half contains zero entries. The set of two matrices E and P is closed relative to multiplication and it defines the known multiplication table of hyperbolic numbers (Figure 6, at the bottom right).

$$M=M_L+M_R=\begin{bmatrix} a_0 & a_1 & -a_3 & -a_2 \\ -a_1 & a_0 & -a_2 & a_3 \\ a_1 & -a_0 & a_2 & -a_3 \\ -a_0 & -a_1 & a_3 & a_2 \end{bmatrix} = E*K+P*K, \text{ where } K=\begin{bmatrix} a_0 & a_1 & -a_3 & -a_2 \\ -a_1 & a_0 & -a_2 & a_3 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}; \quad P=\begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}; \quad P^2=E;$$

	E	P
E	E	P
P	P	E

Figure 6. The decomposition E*K +P*K of the sum $M=M_L+M_R$ of two (4*4)-matrix representations of complex numbers M_L and M_R inside a 4-dimensional space (from Figure 5). Here E is unit matrix and the permutation matrix P is the (4*4)-matrix representation of imaginary unit of hyperbolic numbers with its property $P^2 = E$

The expression M=E*K+P*K resembles the known expression of hyperbolic numbers w=1*x+j*y (where $j^2 = +1$; x and y are real numbers; see details in http://en.wikipedia.org/wiki/Split-complex_number), but with the following differences:

1) the expression M=E*K+P*K includes (4*4)-matrix K as a «coefficient» of the basic elements E and P instead of real numbers x and y in the case of hyperbolic numbers w=1*x+j*y;
2) each of the two summands E*K and P*K is non-commutative product of matrices (permutations of the factors changes the matrix M) instead

commutative product in two summands 1*x and j*y in the case of hyperbolic numbers.

The (4*4)-matrices E and P can be rewrited by means of tensor multiplication ⊗ of (2*2)-matrices: E=[1 0; 0 1]⊗[1 0; 0 1], P=[0 1; 1 0]⊗[0 -1; -1 0], where [1 0; 0 1] is unit matrix of the 2-nd order; [0 1; 1 0] and [0 -1; -1 0] are imaginary units of hyperbolic numbers. It gives the following expression for the sum $M=M_L+M_R$ of two (4*4)-matrix representations of complex numbers M_L and M_R inside a 4-dimensional space:

$$M=E*K+P*K = ([1\ 0;\ 0\ 1]\otimes[1\ 0;\ 0\ 1])*K + ([0\ 1;\ 1\ 0]\otimes[0\ -1;\ -1\ 0])*K \qquad (1)$$

The participation of tensor multiplication ⊗ in the expression (1) allows classifying matrices M as one of members of a set of so called "tensor-numbers" (tensor-complex numbers, tensor-hyperbolic numbers, tensor-quaternions, etc.), which have been revealed and described by the author in his study of genetic matrices and genetic projectors [3]. Multi-dimensional tensor-numbers are a generalization of many kinds of hypercomplex numbers and have specific properties. For example, tensor-complex numbers can be added, subtracted, non-commutative multiply and divided with each other. As the author can judge, tensor-numbers have been never used in mathematical natural science. Consequently the study of bioinformatics has led to new kinds of multi-dimensional numbers (and appropriate multi-dimensional operators) in mathematical natural science. Here one can remind the statement by J.Fourier [7, p.7]: "Profound study of nature is the most fertile source of mathematical discoveries". These new kinds of multi-dimensional numbers (and appropriate operators in spaces of functions), which have been discovered in the field of bioinformatics, have wide perspectives of their applications not only in bioinformatics but also in systems of artificial intellect, robotics, theory of communication, theoretical physics, etc.

After the discovery of non-Euclidean geometries and of Hamilton quaternions, it is known that different natural systems can possess their own geometry and their own algebra [8]. The genetic code is connected with its own multi-dimensional numerical systems or the multi-dimensional algebras. It seems that many difficulties of modern bioinformatics and mathematical biology are connected with utilizing for their natural structures inadequate algebras, which were developed for completely other natural systems.

Pythagoras has formulated the idea: "*all things in the world are numbers*" or "*number rules the world*". Our researches of oblique projectors in the field of matrix genetics have led to new systems of multidimensional numbers and have given new materials to the great idea by Pythagoras in its modernized formulating: "*All things are multi-dimensional numbers*".

Some concluding remarks. This article proposes a new mathematical approach to study "*a partnership between genes and mathematics*" [1]. In the author's opinion, tensor-numbers and operators on their bases give a beautiful mathematical theory, which can be used for further developing of algebraic biology and theoretical physics in accordance with the famous statement by P. Dirac, who taught that a creation of a physical theory must begin with a beautiful mathematical theory: "*If this theory is really beautiful, then it necessarily will appear as a fine model of important physical phenomena. It is necessary to search for these phenomena to develop applications of the beautiful mathematical theory and to interpret them as predictions of new laws of physics*" [9]. According to Dirac, all new physics, including relativistic and quantum, are developing in this way. One can suppose that this statement is also true for mathematical biology.

Materials of this paper reinforce the author's point of view that living matter in its informational fundamentals is an algebraic essence. Projection operators are one of the most useful notions and mathematical instruments to study the genetic coding system and genetically inherited biological phenomena. The author believes that a development of algebraic biology is possible. By analogy with the known fact that molecular foundations of genetics turned up unexpectedly very simple, perhaps algebraic foundations of living matter are also relatively simple.

References

[1] Stewart, I. *Life's other secret: The new mathematics of the living world*. New-York: Penguin, 1999.

[2] Petoukhov S.V. (2008) *Matrix genetics, algebras of the genetic code, noise immunity*. M., RCD, 2008 (in Russian).

[3] Petoukhov, S.V. The genetic code, algebra of projection operators and inherited biological ensembles. - http://arxiv.org/abs/1307.7882 , 2012.

[4] Petoukhov, S.V., He, M. *Symmetrical Analysis Techniques for Genetic Systems and Bioinformatics: Advanced Patterns and Applications*. Hershey, USA: IGI Global, 2009.

[5] Konopelchenko, B. G., Rumer, Yu. B. Classification of the codons in the genetic code. *Doklady Akademii Nauk SSSR, 223(2),* 145-153, 1975 (in Russian)

[6] Rumer, Yu. B. Systematization of the codons of the genetic code. - *Doklady Akademii Nauk SSSR,* vol. 183(1), p. 225-226, 1968 (in Russian).

[7] Fourier, J. *The analytical theory of heat*. – Cambridge: Deighton, Bell and Co, 1878. - https://archive.org/details/analyticaltheory00fourrich

[8] Kline, M. *Mathematics. The loss of certainty*. New York: Oxford University Press, 1980.

[9] Arnold, V. A complexity of the finite sequences of zeros and units and geometry of the finite functional spaces. - *Lecture at the session of the Moscow Mathematical Society*, 2007 http://elementy.ru/lib/430178/430281.

Golden and Harmonic Mean in the Genetic Code

Miloje Rakočević [a]
Department of Chemistry, Faculty of Science, University of Niš, Serbia

Abstract

In previous two works [1, 2] we have shown the determination of the genetic code by golden and harmonic mean within standard Genetic Code Table (GCT), i.e. nucleotide triplet table, whereas in this paper we show the same determination through a specific connection between two tables - of nucleotide doublets Table (DT) and triplets Table (TT), over polarity of amino acids.

1 Introduction

In a previous work we have shown that golden mean is a characteristic determinant of the genetic code (GC), regarding on the codons binary tree, 0-63 [1]. In a second one we showed a splitting of Genetic Code Table (GCT) into three equal and significant parts, using the harmonic mean ($H(a,b) = 2ab/(a+b); a = 63, b = 31.5$) [2]. In this paper, however, we will show that a specific unity of golden mean and harmonic mean appears to be the determinant of Rumer's Table of 16 nucleotide doublets [3] (Tables 1 & 2 in relation to Tables 3 and 4).

As we have shown, golden mean "falls" between the 38th and 39th codon (38. CAA, 39. CAG), which code for glutamine (Q), a more complex of

[a] e-mail address: milemirkov@open.telekom.rs

01. G	GG (6)	**02. F**	UU (4)	**03. L**
04. P	CC (6)	**05. N**	AA (4)	**06. K**
07. R	CG (6)	**08. I**	AU (4)	**09. M**
10. A	GC (6)	**11. Y**	UA (4)	**12. St.**
13. T	**AC** (5)	**14. H**	**CA** (5)	**15. Q**
16. V	GU (5)	**17. C**	UG (5)	**18. W**
19. S	UC (5)	**20. D**	GA (5)	**21. E**
22. L	CU (5)	**23. S**	AG (5)	**24. R**

Table 1. Rumer's Table of nucleotide doublets

01. G	GG (6)	**02. F**	UU (4)	**03. L**
04. P	CC (6)	**05. N**	AA (4)	**06. K**
07. A	GC (6)	**08. Y**	UA (4)	**09. St.**
10. R	CG (6)	**11. I**	AU (4)	**12. M**
13. V	**GU** (5)	**14. C**	**UG** (5)	**15. W**
16. T	AC (5)	**17. H**	CA (5)	**18. Q**
19. L	CU (5)	**20. S**	AG (5)	**21. R**
22. S	UC (5)	**23. D**	GA (5)	**24. E**

Table 2. The modified Rumer's Table

Table 1: Distributions of AAs after nucleotide doublets presented in Table 2: Four squares with dark tones (outer) contain four first doublets from Table 2 and four light (inner) contain four second doublets. In amino acids (within their side chains) at outer/inner areas there are 369/369 nucleons and 61/61 atoms, respectively. All AAs in outer area are nonpolar whereas those in inner area are polar, measured by cloister energy.

only two amide amino acids (AAs); two codons, adjacent to the codons (40.UGU, 41.UGC), which code for one of the only two sulfur AAs, cysteine (C). This "harmonization" of diversity is increased by the harmonic mean, in position 42 on the sequence 0-63. The harmonization extends further to "stop" codon (42.UGA) and to codon (43.UGG) that codes for the most complex AA, tryptophan (W). (The "42" as ending position on the "Golden route" - with Fibonacci numbers - on the Farey tree, corresponding with six-bit GC binary tree [1].)

On the other side, the splitting of GCT into three parts through harmonic mean [2] makes that AAs are distinguished on the basis of the validity of the evident regularities of key factors, such as polarity, hydrophobicity and enzyme-mediated AAs classification (with parameter values as in Table 2.1 in Rakočević, 2013).

2 A new rearrangement of nucleotide doublet Table

With a minimal modification of Rumer's nucleotide doublets Table (DT) follows the next result: if at the beginning of first sub-system[1], with 6/4 hydrogen bonds, are GG/UU doublets, chemical reasons require GU/UG doublets at the beginning of the second sub-system, with 5/5 hydrogen bonds, instead of AC/CA as it is in Table 1. From the same reasons, we have the changes: CG/GC & UC/CU on the left and AU/UA & GA/AG on the right. With the four first doublets we have four outer squares, i.e. codon families (n_1 = GG,UU,GU,UG) in nucleotide triplets Table (TT) which code for nonpolar AAs; the four second doublets give four inner codon families (n_2 = CC,AA,AC,CA), which code for polar AAs (Table 3)[2]. With the four third (n_3 = GC,CU;UA,AG) and four fourth (n_4 = CG,UC;AU,GA) doublets are chosen eight intermediate codon families, the first three code for nonpolar AAs [(CU,AU,GC) → (L, I, M, A)] and the second five for polar AAs [(UC,UA,CG, AG, GA) → (S,Y, R, D, E)] (Table 4). As we see the polar/nonpolar distribution of squares in Table 3 is realized as 4±0 and in Table 4 as 4±1. By this, the polarity/nonpolarity is taken after cloister energy as in Ref. [4][3].

3 Particles number balances through polarity

Distinctions through polarity, presented in Tables 3 & 4, are followed by the balance of the number of nucleons and atoms. Irrespectively of the Table of nucleotide doublets, V. shCherbak showed [Ref. 10, Fig. 10, p. 173] that it makes sense to display the Table of nucleotide triplets (TT) exactly as here in Tables 3 & 4: four squares at the corners and four squares in the center as in Table 3; then, eight squares in middle, i.e. in "between" positions, as in Table 4. After shCherbak's view, the balance of the number of nucleons is the next: AAs in four squares in the corners of TT as well as AAs in four

[1]Two subsystems, each with two quadruplets; in total four quadruplets: two on the left as one-meaning (each nucleotide doublet codes for one AA) and two on the right as two-meaning (each nucleotide doublet codes for two AAs or for one AA and termination in the protein synthesis).

[2]Histidine (H) is neutral in cloister energy with the value ± 0 [4], but polar in hydropathy [5], polar requirement [6, 7] and in hydrophobicity [8, 9].

[3]Cloister energy is "a formal free energy (= cloister energy) of transfer of the amino acid from the outside of a protein to the inside I use cloister energy in preference to other measures of amino acid hydrophobicity-philicity because it is an in situ measure of the property of interest" [4].

1st	2nd letter				3rd
	U	C	A	G	
U	UUU **F** UUC UUA **L** UUG	UCU UCC UCA **S** UCG	UAU **Y** UAC UAA CT UAG	UGU UGC **C** UGA CT UGG **W**	*U* *C* *A* *G*
C	CUU CUC CUA **L** CUG	CCU CCC CCA **P** CCG	CAU **H** CAC CAA **Q** CAG	CGU CGC CGA **R** CGG	*U* *C* *A* *G*
A	AUU AUC **I** AUA **M** AUG	ACU ACC **T** ACA ACG	AAU **N** AAC AAA **K** AAG	AGU AGC **S** AGA AGG **R**	*U* *C* *A* *G*
G	GUU GUC **V** GUA GUG	GCU GCC **A** GCA GCG	GAU **D** GAC GAA **E** GAG	GGU GGC **G** GGA GGG	*U* *C* *A* *G*

1st	2nd letter				3rd
	U	C	A	G	
U	UUU **F** UUC UUA **L** UUG	UCU UCC UCA **S** UCG	UAU **Y** UAC UAA CT UAG	UGU **C** UGC UGA CT UGG **W**	*U* *C* *A* *G*
C	CUU CUC CUA **L** CUG	CCU CCC CCA **P** CCG	CAU **H** CAC CAA **Q** CAG	CGU CGC CGA **R** CGG	*U* *C* *A* *G*
A	AUU **I** AUC AUA **M** AUG	ACU ACC **T** ACA ACG	AAU **N** AAC AAA **K** AAG	AGU **S** AGC AGA **R** AGG	*U* *C* *A* *G*
G	GUU GUC **V** GUA GUG	GCU GCC **A** GCA GCG	GAU **D** GAC GAA **E** GAG	GGU GGC **G** GGA GGG	*U* *C* *A* *G*

Table 3: Distributions of AAs after nucleotide doublets presented in Table 2: Four squares with dark tones (outer) contain four first doublets from Table 2 and four light (inner) contain four second doublets. In amino acids (within their side chains) at outer/inner areas there are 369/369 nucleons and 61/61 atoms, respectively. All AAs in outer area are nonpolar whereas those in inner area are polar, measured by cloister energy.

squares in the center of TT have 369 nucleons [(F91 + L57 + V43 + G01 + W130 + C47 = 369); (P41 + T45 + K72 + N58 + Q72 + H81 = 369)].

To this shCherbak's insight, we now add: the same quantity give the AAs in right site of TT; the right site in relation to the diagonal F-G[4] in TT (S31+Y107+R100+S31+R100 = 369). On the left side of the diagonal there are 336 nucleons (L57+I57+M75+A15+D59+E73=336), what means 33 nucleons less, in relation to 369. With this emergence of difference of "33" on the scene appears a specific self-similarity because the number 33 is an important determinant of the number of atoms in the rows and columns of GCT, i.e. of TT[5].

To this self-similarity determination by nucleon number we now also add the self-similarity determination by atom number: AAs in four squares in the corners as in the center of TT have 61 atoms in AA side chains [(F14+L13+V10+G01+W18+C05=61); (P08+T08+K15+N08+Q11+H11=61)]. In relation to the diagonal F-G, in TT, there are 58 and 59 atoms, respectively; on the left: L13+I13+M11+A04+D07+E10=58, and on the right: S05+Y15+R17+S05+R17=59. These quantities (58 and 59) are the same as the quantities of hydrogen atoms in Sukhodolets' system (what is a further self-similarity): 58 in two inner and 59 hydrogen atoms in two outer rows [Ref. 11, Tab. 7, p. 830], [12]. (Notice that 58 + 59 = 117 is total number of hydrogen atoms in 20 canonical AAs of GC, within their side chains, what is the self-similarity once more.)

4 Determination through "golden whole"

The splitting of GCT (i.e. TT) into 4 outer, 4 inner and 8 intermediate squares, corresponding to responsible nucleotide doublets, leads us to the following conclusion. Within the set of all n-gons, where n is even number, the case n = 4 is only and one case where harmonic mean of "golden whole" $(n^2 - n)$[6] and its

[4]Starting from diagonal (F-G) together with two adjacent ones (S-R and L-E) we give a central space ("in" space) in a strict balance with the set of polar AAs, polar through hydropathy [5]: atom number in two and two sets ("in", "out" / polar, nonpolar) differs exactly for ±1 and ±10, respectively [Ref. 11, Tab. A.3, p. 840 in relation to equations 11-14, p. 838].

[5]If we consider the set of "61" of AAs, then in two rows, YNR & RNY, there are 8 x 33 and in two other, YNY & RNR, 10 x 33 atoms. On the other hand, in two pyrimidine columns, NYN, there are (9 x 33) - 1 and in two purine ones, NRN, (9 x 33) + 1 of atoms [Ref.12, Tab. 3a, p. 224]. (Y for pyrimidine, R for purine and N for all four types of nucleotides.)

[6]The full equation for "golden whole" can have the next form: $n^2 \pm n$ = "WHOLE"; for n = 4, the difference is 12 and the sum 20. In relation to genetic code there are self-similarities expressed through mathematical operations: (4 x 4) - 4 = 12; (4 x 4) + 4 = 20, (4 x 4) x 4 = 64. [Number 12, correspondent to 4+4+4 doublets in sequence $[n_1 - (n_3 \, or \, n_4) - n_2]$; number 20 as 20 AAs (4 x 4 = 16 AAs of alanine stereochemical

half $[(n_2 - n)/2]$ equals $2n$, and $n^2 - n = 3n$. So, in this case we have that the ratio 2:3 appears to be the harmonic mean within the harmonic mean, and, by this, the sequence $n_1 - (n_3\, or\, n_4) - n_2$ corresponds with the Cantorian triadic set. Moreover, such a harmonic mean appears to be corresponding with the number of "small squares" within intermediate space in form of only one "ring" as it follows: [(2 + 2) + (4 x 0) = 4]; [(4 + 4) + (8 x 1) = 16]; [(6+6) + (12 x 2) = 36]; [(8+8) + (16 x 3) = 64] etc. As it is self-evident, the symmetrical "out - middle- in" arrangement (1:1:1 of rings) is not possible for $n \neq 4$, neither for n-gons nor for n-letter alphabets. At the same time here is a self-similarity expressed through the number of "small squares" in the sequence $n_1 - (n_3\, or\, n_4) - n_2$ and the number of codons within them: four squares per n_1, n_2, n_3, n_4, each square per four codons. Moreover, there is a self-similarity between golden and harmonic mean versus 4-letter alphabet: 1n as 1 square (1 nucleotide doublet), 2n as harmonic mean (in the sense above said), $3n$ as golden whole and $4n$ as the sum $n_1+n_2+n_3+n_4$; all these versus 1 letter of alphabet (as letter minimum), 2 letters as word root (nucleotide doublet), 3 letters as 3-letter word (codon) and 4 letters as letter maximum within alphabet[7].

5 Concluding remarks

With the title of the paper is given a working hypothesis that the golden mean (GM) and harmonic mean (HM) are determinants of the genetic code. The findings presented by four illustrations show that this hypothesis is confirmed. However, unlike the previous access to the same determination, it refers not only the analysis of the nucleotide triplet Table, but rather refers to the two tables - Table of doublets (DT) and Table of triplets (TT). In fact, it is precisely presented that these tables are unique in terms of determination just over GM and HM. It is expected that all these uniqueness correspond to the same, or similar uniqueness, found by other authors ([16, 17, 18, 19]), what in future researches should be checked. However, presented facts are such that ones reaffirm the other and vice versa. All together, they favor the recognition that the chemical reactions that determine the GC are not only the reactions in a "test tube", but these reactions are associated with a specific balance of the number of particles (atoms and nucleons); balance, determinated by unique arithmetic and algebraic regularities and expressed in the form of specific (nonfractal) self-similarity ("a harmonized chemistry"). From this it follows further that presented facts also support the hypothesis that the genetic code was from very begining, in prebiotic times and conditions, a complete code [10, 15]. On the other hand, the knowledge that "the chemistry of living" is actually a harmonized chemistry requires great care in medicine, agriculture and natural environment, taking into account the fact that this harmonization is

types and 4 of non-alanine stereochemical type); number 64 as 64 codons.] (About four stereochemical types see Ref. [13] and [14].)

[7]By this one must notice that all these self-similarities are possible only for 4-letter alphabet and 3-letter words.

strictly immanent to the living as such, mediated by genetic code as such.

Acknowledgments

I am very grateful to Branko Dragovich for the invitation to participate in TABIS.2013 conference. Also my thanks belong to him, and then to Vladimir Ajdačić and Tidjani Negadi for help and support in my research, for stimulating discussions and benevolent critique.

References

[1] M. M. Rakočević, "The genetic code as a Golden mean determined system," *Biosystems* **46**, 283–291 (1998).

[2] M. M. Rakočević, "Harmonic mean as a determinant of the genetic code," arXiv:1305.5103v4 [q-bio.OT] (1998).

[3] Yu. B. Rumer, "O sistematizacii kodonov v geneticheskom kode," *Dokl. Akad. Nauk. SSSR* **167**, 1393–1394 (1966).

[4] R. Swanson, "A unifying concept for the amino acid code," *Bull. Math. Biol.* **46**, 187–207 (1984).

[5] J. Kyte and R.F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.* **157**, 105–132 (1982).

[6] C.R. Woese et al., "On the fundamental nature and evolution of the genetic code," in *Cold Spring Harbor Symp. Quant. Biol.* **31**, 723–736 (1966).

[7] B. G. Konopel'chenko and Yu. B. Rumer, "Klassifikaciya kodonov v geneticheskom kode," *Dokl. Akad. Nauk. SSSR* **223**, 471–474 (1975).

[8] S. D Black and D. R Mould, "Development of hydrophobicity parameters to analyze proteins, which bear post- or cotranslational modifications," *Anal. Biochemistry* **193**, 72–82 (1991).

[9] V.R. Chechetkin and V.V. Lobzin, "Stability of the genetic code and optimal parameters of amino acids," *J. Theor. Biol.* **269**, 57–63 (2011).

[10] V. I. Shcherbak, "The arithmetical origin of the genetic code," in: *The Codes of Life* (Springer, 2008).

[11] M. M. Rakočević, "Genetic code as a coherent system," *Neuroquantology* **9** (4), 821–841(2011); (www.rakocevcode.rs)

[12] V. V. Sukhodolets, "A sense of the genetic code: reconstruction of the prebiological evolution stage," *Genetics* **XXI** (10), 1589–1599 (1985) [in Russian].

[13] E. M. Popov, *Strukturnaya Organizaciya Belkov* (Nauka, Moscow, 1989) [in Russian].

[14] M. M. Rakočević and A. Jokić, "Four stereo chemical types of protein amino acids: synchronic determination with chemical characteristics, atom and nucleon number," *J. Theor. Biol.* **183**, 345–349 (1996).

[15] M. M. Rakočević, "A harmonic structure of the genetic code," *J. Theor Biol.* **229**, 463–465 (2004).

[16] B. Dragovich, "p-Adic structure of the genetic code," *NeuroQuantology* **9** (4), 716–727 (2011); arXiv:1202.2353v1 [q-bio.OT].

[17] T. N gadi, The Multiplet structure of the genetic code, from one and small number, *Neuroquantology* **9** (4), 767–771 (2011).

[18] N. Ž. Mišić, "Nested numeric/geometric/arithmetic properties of shCherbak's prime quantum 037 as a base of (biological) coding/computing," *Neuroquantology* **9** (4), 702–715 (2011).

[19] F. Castro-Chavez, "The quantum workings of the rotating 64-grid genetic code," *Neuroquantology* **9** (4), 728–746 (2011).

The Structure of Emotional Dialogs in Online Social Networks: High-Arousal Clustering

Bosiljka Tadić [a]
Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, Slovenia

Vladimir Gligorijević [b]
Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, Slovenia

Milovan Šuvakov [c]
Institute of Physics, University of Belgrade, Belgrade, Serbia

Abstract

Recent advancements in psychology as well as in new fields of affective computing make possible quantitative study of emotion at various levels of human behavior. In this work, we analyse the collective emotional response in online social networks that can arise in self-organized dynamics from individual activity of many

[a] e-mail address: bosiljka.tadic@ijs.si
[b] e-mail address: vgligorijevic@gmail.com; current address:Department of Computing, Imperial College London, London SW7 2AZ, UK
[c] e-mail address: suvakov@gmail.com

interacting users. In online social network MySpace, the appearance of collective dynamics is shown by computing the avalanches of dialogs which carry emotional contents. Moreover, by adopting agent-based simulations we explore the salient features of the underlying self-organized dynamics; our focus is on events with high emotional arousal, which triggers agents' activity. In the model, the rules and the agents attributes as well as the network structure are inferred from the empirical data of MySpace. Quantitative analysis of the simulated data reveals how the collective dynamics in online social networks is organized based on the extrinsic driving noise.

1 Introduction

In human behavior emotions [1, 2] play an important role from the level of brain activity of each individual to social interactions, and large-scale social phenomena [3, 4]. Recent developments in the quantitative study of emotion based on Russell model in psychology [5] and machine-learning methods of text interpretation [6], emotion components (*arousal*–degree of reactivity, and *valence*–degree of pleasure or displeasure) can be retrieved from text messages in online communications among users [7]. Quantitative science of human dynamics on the Web in the framework of statistical physics of complex systems, supplied with the machine-learning methods of text analysis, strives to reveal the role of emotions in collective behaviors of users in the virtual world [8]. Applying these approaches for the empirical data analysis and agent-based modeling, we have studied several online communication systems (a survey is available online [9]).

Online Social Networks (OSN), such as `MySpace` and `Facebook`, represent a specific type of virtual networks in which communications follow "friendship" links, closely resembling off-line social systems [10, 11]. However, a detailed analysis of the dynamics in OSN suggest that they possess own regularities and different structure of connections [10]. These are related with the altered role that an individual assumes in the virtual world [12, 13] as well as with the use of emotions [14]. The analysis of empirical data from `MySpace` social network [10] indicates that, by various mechanisms in the dynamics, the emotional dialogs can lead to bursting events that involve many users; moreover, the emergence of specific patterns of connections, e.g. starlike structures with a strong central node, is dynamically realized by different use of the available social links.

In this paper, we focus on the stochastic dynamics of emotional commu-

nications, which enables the occurrence of collective emotional behaviors in OSN. First, with the analysis of empirical data from MySpace dialogs, we provide evidence that the emotional clustering of events occurs in OSN. Then, keeping the network structure as well as the rules and parameters of the empirical system, we perform agent-based simulations of the emotional dialogs between agents. In the analysis of the simulated data, we focus on the events carrying high emotional arousal, which is mainly responsible for the activity of agents. Fractal analysis of the corresponding time series and bursting events (avalanches) enables us to quantify the underlying stochastic processes and helps to understand the collective dynamics in these networks.

2 Evidence of collective dynamics in OSN

2.1 Structure of the empirical network

The empirical data that we consider here have been collected and studied in [10]; the data consist of the networked dialogs in MySpace occurring within a fixed temporal depth of two months. Each message from user i to user j registered in the considered time window is represented by a directed link $i \to j$. Multiple messages along the same link contribute to the width of the directed link. The network of dialogs has a characteristic structure with a large number of communities and broad distributions of in-coming and outgoing degree and widths [10]. However, its pronounced disassortativity [10] is in a strong disagreement with the conventional off-line social networks. A close-up part of the network structure (inside of a community) is shown in Fig. 1. A dominant red color of the links indicates that positive emotion as it was detected by the emotion classifier [15] in the texts of each message, prevails in all messages exchanged along these links.

2.2 Avalanches of emotional messages in MySpace social network

The underlying dynamics of dialogs can be studied in detail considering time series of the number of messages and messages which carry specified emotional content [10]. Clustering of event (avalanches) can be determined from the same time series. Specifically, an avalanche consists of the part of time series, which is above a base line between two consecutive intersections with the base line [16, 17]. Considering the time series of messages which carry positive and negative valence, we determine avalanches of such

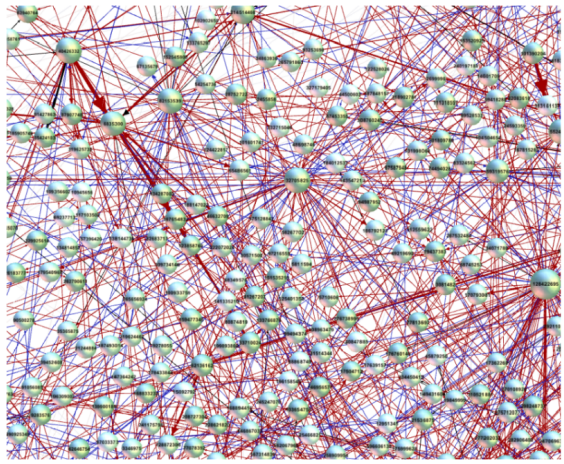


Figure 1: A part of social network `MySpace` with users, shown as nodes, exchanging emotional messages along the "friendship" links.

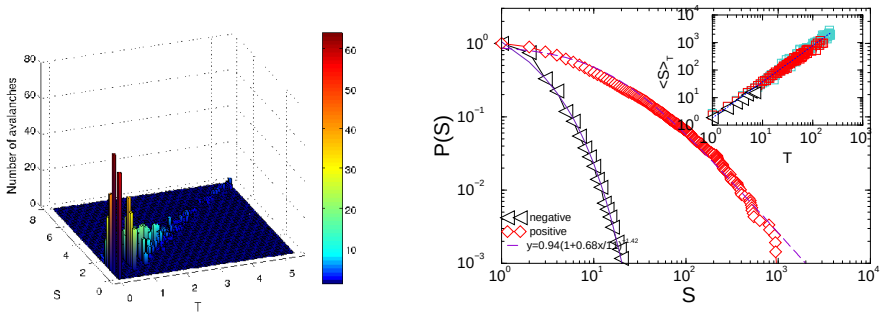


Figure 2: Left: 3D histogram $P(S,T)$ of the number of avalanches of size S and duration T. Right: The distribution of avalanche sizes for positive/negative emotion avalanches (main) and the avalanches geometry factor (inset).

messages in the empirical data. The distributions of avalanche sizes are shown in Fig. 2 (right). In the inset we plot the average size $< S >_T$ of all avalanches which have a fixed duration T. The power-law dependence $< S >_T \sim T^{\gamma_{ST}}$ is in agreement with the scaling of avalanches in self-organized dynamics [18, 19]; the exponent $\gamma_{ST} = 1.25$ measures the geometry (spread) of avalanches. The avalanches of all messages, independently on their emotional contents, follows the same scaling law. As it is obtained from the same empirical dataset, the 3-dimensional histogram of such avalanches is shown in Fig. 2(left).

3 Theoretical study within agent-directed simulations

To understand mechanisms of emotional bursts, we use agent-based modeling approach and perform simulations of agent's interaction along the links of the same social network `MySpace` that is discussed above in Sect. 2. In the model, the users in the OSN are represented by *agents*, whose properties are related to the corresponding features of users in the same empirical data. According to recent advances in agent-based modeling approach to social dynamics [20], it is essential to define the agent's profile by specifying the following attributes:

$$A\{id, (a_i(t), v_i(t)), links \in OSN; circadian.cycles; action.delay\} \ . \quad (1)$$

In particular, like in the case of real users, each agent has an unique *id*, given by the number $i = 1, 2 \cdots N$, N is the network size; the agent's personal social connections in the network are specified by the list of all links; the emotional state of each agent is defined by two state variables, arousal $a_i(t)$, and valence $v_i(t)$, which fluctuate in time under the influences that an agent receives along its social links and the outside world (see below). The *action delay* (or intercativity time) is another characteristic feature of human activity that needs to be taken into account [20]. In the model, after each completed action, the delay time Δt of an agent is taken from the distribution $P(\Delta t)$; the distribution statistically matches the delay times of users in the related empirical system [21, 22, 23]. $P(\Delta t)$ that is determined from the same empirical data is shown in Fig. 3(right). The circadian cycles—daily fluctuations of the agent's activity level—is also taken from the empirical data via the time series $p(t)$, see Fig. 3(right). This time series, representing the number of new arrivals per time step, is used as the *driving force* in the simulations.

The update rules of the model match the rules in the empirical network: the messages are received along personal links; in addition, each user (agent) can see the messages at the "friend's wall" as well as be under a direct influence from the outside world. The rules are schematically depicted in Fig. 3(left). Consequently, the emotional state of an agent i fluctuates according to two nonlinear maps [21]

$$a_i(t+1) = (1-\gamma_a)a_i(t) + \delta_{t_i,1}[\epsilon h_i^a(t) + (1-\epsilon)\bar{h}_i^a(t)] \times [1 - a_i(t)] \ ; \quad (2)$$

$$v_i(t+1) = (1-\gamma_v)v_i(t) + \delta_{t_i,1}[h_i^v(t)] \times [c_1 + c_2(v_i(t) - v_i^3(t))][1 - |v_i(t)|] \quad (3)$$

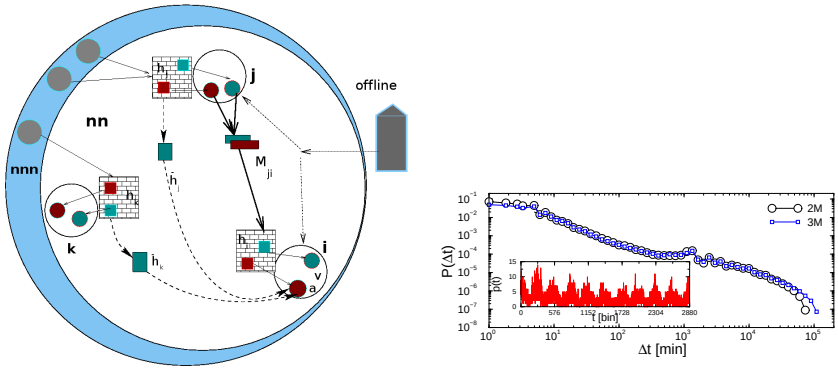


Figure 3: (left) Schematic, building of the agent's i arousal a_i and valence v_i of direct messages from neighbours, viewing walls of next-neighbours and the input from the outside world (Fig. from [21],©BT). (right) Inputs for the model: Time series of new users $p(t)$ and delay-time distribution $P(\Delta t)$ inferred from the empirical data of Ref. [10].

under the influence fields

$$h_i^z(t) = \frac{\sum_j \sum_{m\in M_{ji}} \theta(t,t_m) z_j(t_m) W_{ji} e^{-\gamma^h (t_{ji}^{lm} - t_m)}}{\sum_j \sum_{m\in M_{ji}} \theta(t,t_m) W_{ji} e^{-\gamma^h (t_{ji}^{lm} - t_m)}} e^{-\gamma^h (t - t_{ji}^{lm})} \qquad (4)$$

where the superscript $z = (a, v)$ stands for the arousal and valence field, respectively. In Eq. (4), each message in the stream of messages M_{ji} along weighted W_{ji} personal link $j \to i$ in the network, contributes to build the arousal and valence field of agent i; these fields also decay in time with the rate γ_h. In the arousal dynamics, the additional contribution in Eq. (2), which is balanced by a prefactor $1 - \epsilon$, is due to the messages seen by the agent i at its neighbors' walls, i.e.

$$\bar{h}_i^a(t) = \frac{\sum_j W_{ij} h_j^a(t)(1 + h_j^v(t) v_i(t))}{\sum_j W_{ij}(1 + h_j^v(t) v_i(t))} \ . \qquad (5)$$

In the present simulations, the parameters $\epsilon = 0.9$ is used, assuming a larger weight to received personal messages in comparison with the messages created by a third party at the "friend's wall". For simplicity, the same ratio ϵ applies between two possible actions: replying to a sender of a personal message versus writing to another friend. A detailed description of the model rules and numerical implementation of the program code can be found in Refs. [21]. The external input on the agent's dynamics occurring with a small probability $p_0 = 10^{-2}$, is implemented as a *sudden reset of the emotional states of currently active agents* to either a random value (noisy

input), or to a specified emotion (coherent input). As the coherent input, we simulate two situations with a negative emotion "shame" and a positive emotion "enthusiastic". Apart from the numerical values, in psychology, these emotions differ considerably by their social dimension.

The simulated dynamics—the messages being communicated among agents in the network—results in a temporal stream of messages carrying emotional contents (arousal and valence), in a full analogy to the empirical data studied in Ref. [10]. From the simulated stream of messages, we construct time series of the number of messages with a specified content. For instance, the time series of messages with only positive (negative) emotion valence are particularly interesting in situations where the external inputs are emotionally biased [21].

4 Fractal analysis of the simulated time series

Here, we construct the time series of messages with a *high arousal*. Having in mind that high arousal leads to action of an agent, these messages are particularly interesting from the point of view of the analysis of triggering mechanisms. For this purpose, the messages with high arousal $a > a_{cut}$ are

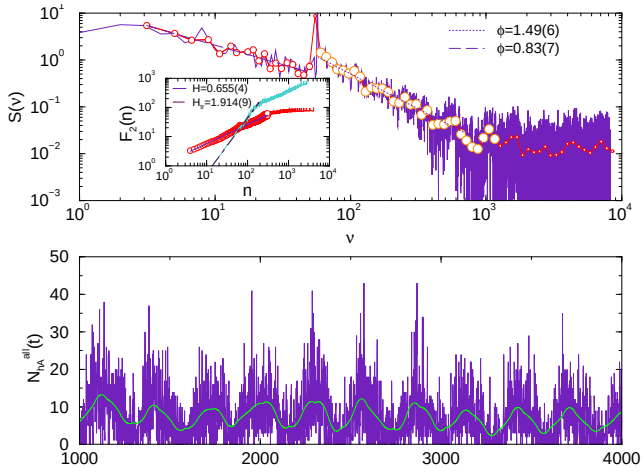
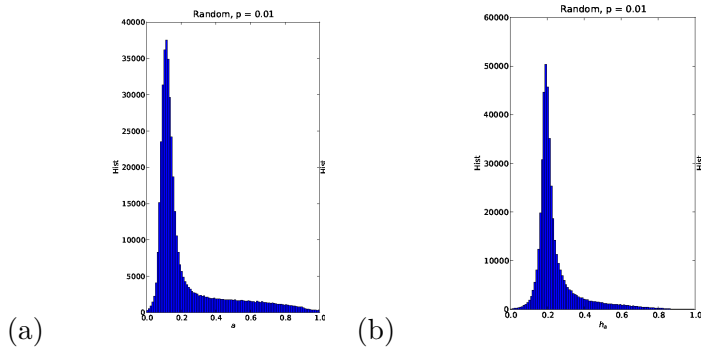


Figure 4: Bottom: Time series of messages with a high arousal $a > a_{cut}$ and trend with daily cycle. Top: Power-spectrum $S(\nu)$ of the time series and (inset) fluctuations $F_2(n)$ around the trend at time interval n of the detrended time series and the trend.

selected. Here, $a_{cut} = 0.23$ in Fig. 5a is the point separating the tail from the Gaussian part of the distribution of arousal of all messages. Time series of the number of messages with the high arousal is shown in Fig. 4. As

expected, these time series also exhibit a pronounced daily cycle (the cycle is depicted by the thick line over the noisy signal). In order to determine the fractal structure of these time series, we firstly remove the cycle. Applying the *detrended time series analysis* after removed cycle (the methodology is described in [10]), we find the scaling region of the fluctuations around the cycle. In the inset to Fig. 4, the scaling region is indicated by the straight line, where the Hurst exponent can be determined. Note that, due to removal of the cycle, the remaining scaling region is shorter than one day. In the scaling region, the fluctuations obey a power-law $F_2(n) \sim n^H$, with the Hurst exponent $H = 0.655 \pm 0.004$. The trend makes almost a perfect cycle, i.e. its Hurst exponent is found $H_{tr} = 1.914 \pm 0.009$. The power spectrum of the original time series is shown in the top panel of Fig. 4. It exhibits two distinct regions where a power-law dependence $S(\nu) \sim \nu^{-\phi}$ can be identified, namely, $\phi = 1.42 \pm 0.06$ in the high-frequency region (corresponding to the interval roughly from one hour to one day in the time domain), and $\phi = 0.83 \pm 0.07$, in the low-frequency domain (time longer than one day). The pronounced peak in the spectrum corresponds to the daily cycle. The observed scaling features suggest differences in the diffusion processes of high-arousal messages within one day time window, as compared with stochastic point processes at a longer time scale.



(a) (b)

Figure 5: Histogram of (a) arousal a and (b) arousal field f^a at the moment of message creation inferred from all messages by agents in the network.

In analogy with *avalanches of events* in the empirical data, shown in Fig. 2, here we determine clustered events with high-arousal messages. The distribution of avalance size $P(s)$ is shown in Fig. 6 for three different external inputs. In each case, the threshold point is determined considering the histograms of arousals as in Fig. 5. The fit curves correspond to q-

exponential [25] distribution $P(s) = A\left(1-(1-q)\frac{s}{s_0}\right)^{1/1-q}$ with different parameters. In the case of random input, the fitted slope of the tail is ~ 1.5, corresponding to $q \sim 1.66$, in a good agreement with $q = 1.68$ which is obtained in the empirical data Fig. 2. On the other hand, in the case of coherent inputs, the slopes are much larger (cf. Fig. 6), leading to a reduced value of the non-extensivity parameter $q \sim 1.33$.

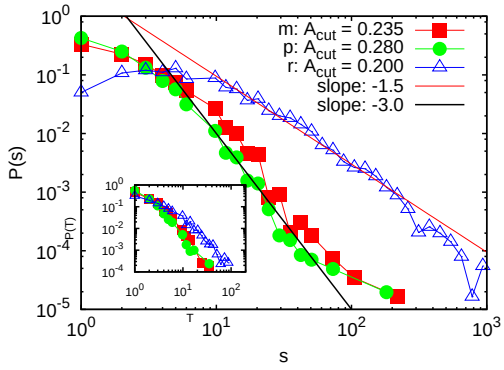


Figure 6: Distribution of avalanche size s (main figure) and duration T (inset) from the high-arousal time series. Different curves correspond to the situations with random input (prefix r:) and two inputs from positive (p:) and negative (m:) emotion valence.

The statistical analysis of *returns*—differences between consecutive values of the dynamic variable—gives valuable information about the nature of the stochastic process [24, 25]. In the present study, the distribution of returns in time series of all messages with high-arousal gives information about the response of the system to the triggering fields. In Fig. 7 the distributions corresponding to the random and coherent inputs are depicted. The fits with q-Gaussian [24, 25] expression $P(d) = A\left[1-(1-q)\left(\frac{d}{d_0}\right)^2\right]^{1/1-q}$ with $q = 1.45$ and $q = 1.66$, respectively, suggest that different relaxation mechanisms occur in the case of random compared with the coherent noise.

5 Conclusions

The agent-based simulations, here employed in OSN, reveal how the collective emotional behavior can arise in the stochastic process where emotional messages are communicated between participants. In resemblance to the original empirical data, the results derived from simulated dynamics—persistence, correlations and clustering of events that convey a spec-

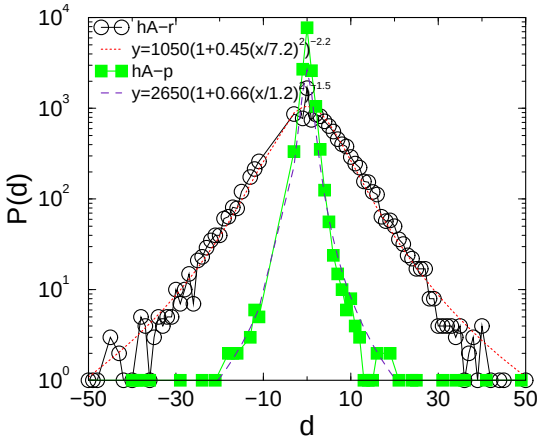


Figure 7: Histogram of returns from high-arousal time series with random (circles) and a coherent input with the emotion "enthusiastic" (squares). Fits according to q-Gaussian distribution with parameters: $q = 1.45$, $d_0 = 7.2$, for random, and $q = 1.66$, $d_0 = 1.2$, for coherent input.

ified emotional content—are compatible with non-extensive Tsallis statistics. Particular attention was given to events with high-arousal messages. The associated relaxation processes in the case of random noise inputs are in a good agreement with the empirical system `MySpace`. However, the quantitative measures are quite different when coherent noise with a picked emotion ("enthusiastic") is simulated. It is interesting to note that, in the latter case, the relaxation non-extensivity parameter is comparable to the one measured in the Brain epilepsy dynamics where $q = 1.64$ [24]. Apparently, additional research in the emotion dynamics is needed to span the gap between Brain functions of interacting individuals and collective emotional reactions.

Acknowledgements

We are grateful for support from program P1-0044 by the Research Agency of the Republic of Slovenia and from the project FP7-ICT-2008-3 under grant no 231323. B.T. also thanks for partial support from COST Action KNOWeSCAPE. M.Š. would also like to thank for support from projects OI171037 and III41011 of the Republic of Serbia.

References

[1] J. A. Coan and J. J. B. Allen, editors. *The Handbook of Emotion Elicitation and Assessment*. Oxford University Press Series in Affective Science, 2007.

[2] K. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005.

[3] Ch. von Scheve and M. Salmela, editors. *Collective emotions*. Oxford University Press, 2014.

[4] J. Harding and E.D. Pribram, editors. *EMOTIONS a cultural studies reader*. Rutledge, Oxon, UK, 2009.

[5] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178, 1980.

[6] R.A. Calvo and S. D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18 –37, 2010.

[7] G. Paltoglou, M. Theunis, A. Kappas, and M. Thelwall. Prediction of valence and arousal in forum discussions. *Journal of IEEE Transactions on Affective Computing*, pages 1–9, 2011.

[8] Collective emotions in cyberspace: http://www.cyberemotions.eu/

[9] B. Tadić, M. Mitrović, M. Šuvakov, and V. Gligorijević, Cyberemotions summary: http://www-f1.ijs.si/∼tadic/projects/cybere_.html.

[10] M. Šuvakov, M. Mitrović, V. Gligorijević, and B. Tadić. How the online social networks are used: dialogues-based structure of myspace. *Journal of the Royal Society Interface*, 10(79):20120819, 2012.

[11] E. Ferrara, P. De Meo, G. Fiumara, and A. Provetti. The role of strong and weak ties in facebook: a community structure perspective. *Procedia Computer Science: International Conference on Computational Science, ICCS 2012*, pages 1–10, 2012.

[12] T. Ryan and S. Xsenos. Who uses facebook? an investigation into the relationship between the big five, shyness, narcissism, loneliness, and facebook usage. *Computers in Human Behavior*, 2011.

[13] Y. Amichai-Hamburger and G. Vinitzky. Social network use and personality. *Computers in Human Behavior*, 26(6):1289 – 1295, 2010. Online Interactivity: Role of Technology in Behavior Change.

[14] B. Tadić, V. Gligorijević, M. Mitrović, and M. Šuvakov. Co-evolutionary mechanisms of emotional bursts in online social dynamics and networks. *Entropy*, 15(12):5084–5120, 2013.

[15] G. Paltoglou, M. Thelwall, and K. Buckely. Online textual communcation annotated with grades of emotion strength. In *Proceedings of the Third International Workshop on EMOTION (satellite of LREC): Corpora for research on emotion and affect*, pages 25–30, 2010.

[16] B. Tadić. Nonuniversal scaling behavior of barkhausen noise. *Phys. Rev. Lett.*, 77:3843–3846, 1996.

[17] M. Mitrović, G. Paltoglou, and B. Tadić. Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02005+, 2011.

[18] D. Dhar. Theoretical studies of self-organized criticality. *Physica A*, 369:29, 2006.

[19] Á. Corral. Long-term clustering, scaling, and universality in the temporal occurrence of earthquakes. *Phys. Rev. Lett.*, 92:108501, 2004.

[20] B. Tadić. *Modeling behavior of Web users as agents with reason and sentiment*, pp. 177–186, in "Advances in Computational Modeling Research: Theory, Developments and Applications", A.B. Kora (Ed.) Novapublishing, N.Y., 2013.

[21] M. Šuvakov, D. Garcia, F. Shweitzer, and B. Tadić. Agent-based simulations of emotion spreading in online social networks. http://arxiv.org/abs/1205.6278. 2012.

[22] M. Mitrović and B. Tadić. Dynamics of bloggers' communities: Bipartite networks from empirical data and agent-based modeling. *Physica A*, 391(21):5264 – 5278, 2012.

[23] B. Tadić and M. Šuvakov. Can human-like bots control collective mood: Agent-based simulations of online chats. *Journal of Statistical Mechanics: Theory and Experiment*, P10014, 2013.

[24] G.P. Pavlos, M.N. Xsenakis, L.P. Karakatsanis, A.C. Iliopoulos, A.E.G. Pavlos, D.V. Sarafopoulos. Universality of Tsallis non-extensive statistics and fractal dynamics of complex systems. 2014.

[25] Constantino Tsallis. The nonadditive entropy sq and its applications in physics and elsewhere: Some remarks. *Entropy*, 13(10):1765–1804, 2011.

Self-organizing Internal Representations of Space

Eugenio Urdapilleta [a]
SISSA, Neuroscience, Trieste, Italy

Alessandro Treves [b]
SISSA, Neuroscience, Trieste, Italy

Abstract

Where in the brain is outside space represented? In the mammalian brain, quite clearly it is represented multiple times, and with some differences from species to species. A remarkable representation, however, is that by grid cells, discovered in 2005 in rodents, in the medial entorhinal cortex. Subsequently conjunctive grid-by-head-direction cells were also found, and a similar spatial representation has been observed in bats. This contribution briefly reviews its main features, and argues that it is likely a representation that self-organizes. If so, it is expected that its main properties will reflect those of the environment where the animal is raised.

[a] e-mail address: urdapile@sissa.it
[b] e-mail address: ale@sissa.it

1 Space in the brain: findings in medial Entorhinal Cortex

Spatial cognition and spatial memory are fundamental for any advanced species to survive and thrive in its environment - they facilitate the retrieval of food, escape from danger, and social interaction. One brain structure involved in spatial memory, in mammals, is the so called medial temporal lobe, which includes the hippocampus, the amygdala and related areas of cortex. Place cells in the rat hippocampus, discovered in the early 70's, show elevated firing activity whenever the rat enters a specific portion of the environment, the place field [1]. Head direction (HD) cells in the rat postsubiculum, first observed many years later, are characterized by steady firing when the animal points its head towards a specific direction in the environment [2] - a seemingly simpler representation that requires less detailed memory, if any at all.

In 2005, place-modulated cells were discovered also in the medial entorhinal cortex (mEC), a region just one synapse upstream from the hippocampus. They were dubbed *grid* cells and later observed also in bats [3, 4]. The multiple firing fields of a layer II mEC grid cell collectively form a remarkably regular triangular grid spanning the environment which the animal explores [3]. The grids of different cells, which are close to each other in the neural tissue, share the same spacing and orientation, though they differ in spatial phase. Spatial phases appear to be randomly scattered across neighboring cells, with no relation to their arrangement in the tissue. Thus, whereas place cells can be construed to arise out of relatively simple sparsification mechanisms [5, 6], and head-direction cells could be even thought to be hardwired [7], grid cells seem to require some clever engineering design, that generates their (locally) common periodicity while keeping them distinct in terms of spatial phase. Moreover, a year later, conjunctive grid-by-head-direction cells were observed [8], in medial entorhinal cortex as well, suggesting that these different types of units (all pyramidal cells at the morphological level) should be understood together as components of an integrated system, see Figure 1. How such a system may be set up, however, and what functions exactly it may subserve, has largely remained a mystery.

While grids and conjunctive grid-by-HD units appear to come together, as part of the same endowment, the distinct layers of medial entorhinal cortex (mEC) contain different proportions of cells with specific selectivity, connectivity, and cellular properties. In layer II, in fact, all grid cells found are purely positional, although a significant proportion of cells are not spatially tuned at all. In layer V, the majority of those that do show grid cell properties are conjunctive cells. In layer III and VI, there is a mixture of pure grid cells and conjunctive cells [8, 10]. In several studies, we and others have tried to model the way grid and conjunctive cells express their unique firing patterns, how they may emerge upon learning to move in a new environment or during the development of the animal, and the possible reason for the differentiation between grid and conjunctive cells [11].

A series of modeling studies have proposed mechanisms purely for the *expression* of grid firing patterns [12, 13], without delving into the issue of their *induction*, i.e., how these patterns may be generated initially. In these models, a crucial hypothesis is that the medial temporal lobe spatial representation system is somehow involved in path-integration, i.e. in the calculation of the distance and direction traveled by the animal from a reference point. Grid cells may then express the result of such calculations, or perform them themselves. Path integration may in principle extend over long times, although in practice in relation to grid cells it is thought to bridge over periods of weakened sensory input in the order of seconds, following which sensory cues are available again and reset the path integrator. In path-integration-based models, units accumulate the velocity of the animal to update an estimate of current location, either by the collective state of many units in continuous attractor network models [14, 15, 16, 17], or by the phases of velocity-controlled oscillators at the single unit level, in oscillatory interference models [18, 19, 20, 21]. In both these classes of models the induction of the grid pattern remains unexplained: the triangular grid is imposed *ab initio* by structured collateral connections or by the summation of multiple oscillators with preferred running directions separated by multiples of 60 degrees.

2 A self-organizing model of grid pattern formation

In the self-organization perspective [22], on the other hand, the spatial responses should first emerge spontaneously, at the single unit level. Then, once the grid patterns have emerged and have been wired together through

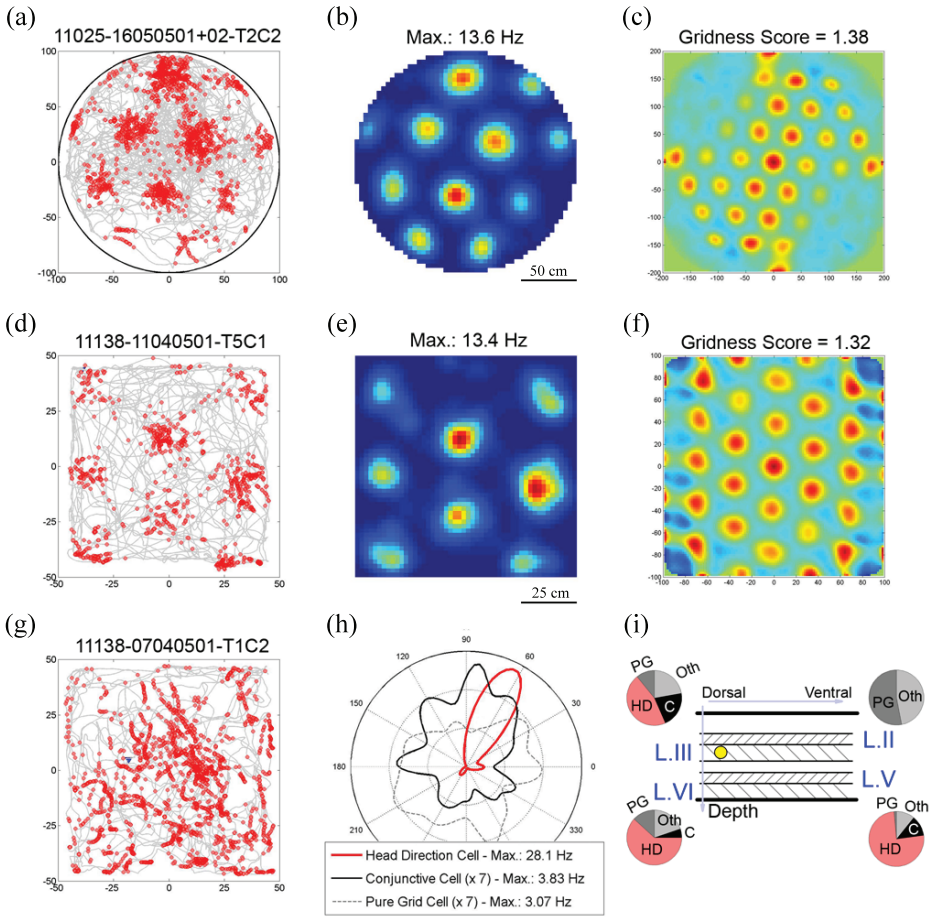


Figure 1: Different functional types of cells found in medial entorhinal cortex (mEC). As a rat explores a small environment, see gray light lines in left panels (a, d, g), spikes are produced in certain spatial regions and/or with certain head directions (superimposed red dots). Top panels (a, b, c) exemplify a "pure grid cell", where field locations are regularly spaced in a hexagonal pattern [see (b) and (c), its autocorrelogram]. As shown by the dashed line in figure (h), this cell does not exhibit head direction selectivity. Intermediate panels (d, e, f) show the typical behavior of "conjunctive cells", which develop the same kind of regular representations [see (e) and (f)], while maintaining an angular preference (see panel h, black line). Figure (g) shows the spiking activity of a "head direction cell" in mEC. No clear spatial association is developed, while head direction is strongly and clearly represented [see red line in (h)]. Figure (i) represents a sketch of the lamination and of the functional differentiation of the cells in mEC. Data was processed from the Sargolini dataset, kindly free-distributed by the Moser lab [9].

recurrent connections, the units that manifest them may also be involved in path-integration. In the simplest version of the model, that we have explored in a series of studies [23, 24], the periodicity of the grid pattern is a result of firing rate adaptation during exploration sessions that span a considerable developmental time. It is fixated gradually by means of synaptic plasticity in the feedforward connections, which convey broad spatial inputs, for example but not necessarily from "place units" [23]. In a recent variant considered by another group, grid units receive inputs from putative periodic "stripe cells" [25], and they inherit the periodicity of these one-dimensional stripe-shaped inputs.

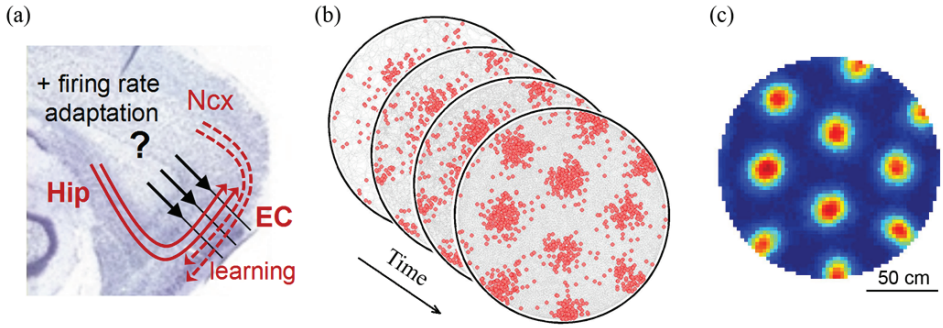


Figure 2: Left: In our self-organized model for the development of grid cells, generic spatially modulated signals coming from the neocortex (Ncx) and the hippocampus (Hip) are merged together in the entorhinal cortex (EC), where learning and firing rate adaptation produce grids. Middle: A regular spatial pattern of spiking emerges from the association between learning and exploration, and from single unit (adaptation) properties. Right: the multi-peaked firing field, with a clear hexagonal structure.

Most models describe either grid units or conjunctive units or are compatible with both, and sometimes critically depend on a feature of either cell type [26], but they do not really relate to the striking phenomenon that both cell types are present, nor do they try to explain how their differential properties may emerge. This differential laminar localization of grid and conjunctive cells in mEC has been the focus of a recent study of ours [11] based on a modified self-organizing adaptation network, in which also the recurrent collateral connections, assumed to exist between conjunctive units, self-organize their weights; while grid units are assumed to be those not endowed with recurrent connections. The analysis of the model,

through extensive computer simulations, indicates a complex time course for self-organization and differential information-theoretic properties of the two populations.

Both grid units and conjunctive units, in the network model, receive inputs from place units. This assumption is consistent with experimental observations: one study [27] shows that inputs from the hippocampus are necessary for grid cells to maintain their grid firing pattern; and during postnatal development, place cells form adult-like spatial fields earlier than grid cells do [28, 29]. Conjunctive units are also interconnected through collaterals, while grid units receive connections from conjunctive units, but have no collaterals among themselves, consistent with recent evidence [30].

In the model, grid units and conjunctive units develop, with time, grid fields resulting from firing rate adaptation and competitive learning. Grid alignment, as shown below, is delayed in both layers with respect to the formation of triangular grids. A common grid orientation among conjunctive units is produced, in the model, by head-direction modulated collateral interactions, while the grids of grid units inherit the same orientation through connections from conjunctive units. Grid units as well as conjunctive units share a similar spacing but show a random distribution of spatial phases. Grid units however carry more spatial information than conjunctive units, thus providing better inputs for the hippocampus to form spatial memories.

2.1 Time scales for grid development

In early development, the firing maps of all the units in the network have multiple fields at irregular locations (see Fig. 3.a, top-left panel). Conjunctive units develop reasonably regular grids, and at the end of the simulation, most conjunctive units in the network develop grids with similar spacing and orientations (see Fig. 3.a, upper panels, for a single unit evolution). The spatial responses of the units in the grid layer also evolve into triangular grids, as synaptic modification proceeds on the connections from place units to grid units and from conjunctive to grid units. The activity of the grid units does not show any preference in head direction, due to the absence of head direction modulation. The grids of the grid units share the same spacing and orientation as those of the conjunctive units, and they are randomly shifted relative to each other, with distributed spatial phases [11].

Since they depend on the aligning input from the conjunctive layer, do grid units develop grids later than conjunctive units? By quantifying the gridness, grid alignment and grid spacing of both layers during develop-

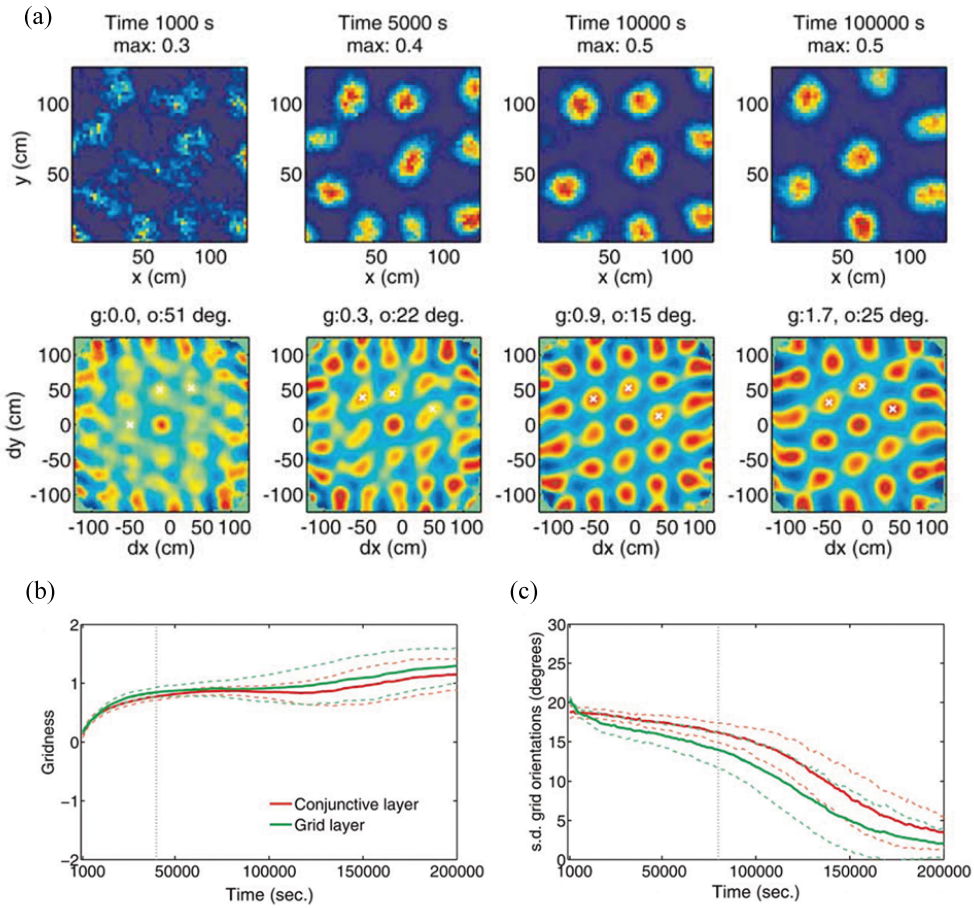


Figure 3: (a) Snapshots of the developing firing fields in a single unit, in the model. Upper panels show the firing fields as time goes. Lower panels present the corresponding autocorrelograms. (b, c) Temporal evolution of the gridness score (b) and the angular dispersion of the orientation (c) in both the conjunctive and the grid layer, during development. Adapted with permission from [11].

ment, we find that averaged across many simulations, the mean gridness scores of the units in both layers increase with a similar time course (see Fig. 3.b). At $4{\times}10^4$ seconds, both layers already form grids with mean gridness scores as high as 0.8. However, mutual grid alignment is reached much later. The standard deviation in grid orientation of the grid units is

below 5 degrees at 1.4×10^5 seconds (38.9 hours, see Fig. 3.c). The conjunctive layer aligns grids on a similar time scale. For conjunctive units, the standard deviation in grid orientation is below 5 degrees at about 1.7×10^5 seconds (47.2 hours). The spacing of the grids in both layers do not change during development. This is not surprising since the adaptation constants do not change during development.

3 Grid development out of the plane

If the model described above captures the essential logic of the differentiation between conjunctive and grid units, assigning to the former the role of enforcing alignment and to the latter that of producing accurate position codes, it also indicates how to proceed in order to address other issues related to the phenomenology of grid firing pattern. It indicates, in fact, that simulating both populations may not be necessary, at least in a first instance, because the pure grids develop *s*laved, in a sense, to the conjunctive units. So understanding the development of conjunctive units seems sufficient.

In a few very recent studies, we have explored the development of conjunctive grid units, in our model, in non-standard environments. These simulation studies are motivated by experiments being undertaken by colleagues.

In the laboratory of Edvard and May-Britt Moser in Trondheim, rats have been raised in a spherical environment, with the aim of observing the grid units that may emerge in such a spatial context, which comprises the entire developmental experience of these animals. Note that in other labs rats are tested, i.e., single units in their medial entorhinal cortex are recorded, while they run on top of a revolving sphere. Mathematically the outside and the inside of the sphere may be equivalent, but a crucial difference is that it does not appear feasible to raise rats on the outside surface of a sphere. What sort of grid cells do we expect to see in rodents who have spent their developmental period inside a large spherical cage? Or, were the alternative experimental paradigm feasible, on a revolving ball, with virtual reality simulating a coherently revolving surround? Our model based on firing rate adaptation predicts that whether experienced on the outside or inside, a spherical environment induces one of a succession of grid maps realized as combinations of spherical harmonics, depending on the relation of the radius to the preferred grid spacing, itself related to the parameters of firing rate adaptation. Numerical simulations concur with the analytical predictions shown in Fig. 4.a. Depending on the radius

of the sphere, one may observe grids with five nearest neighbours to each peak, or even fewer than five [31].

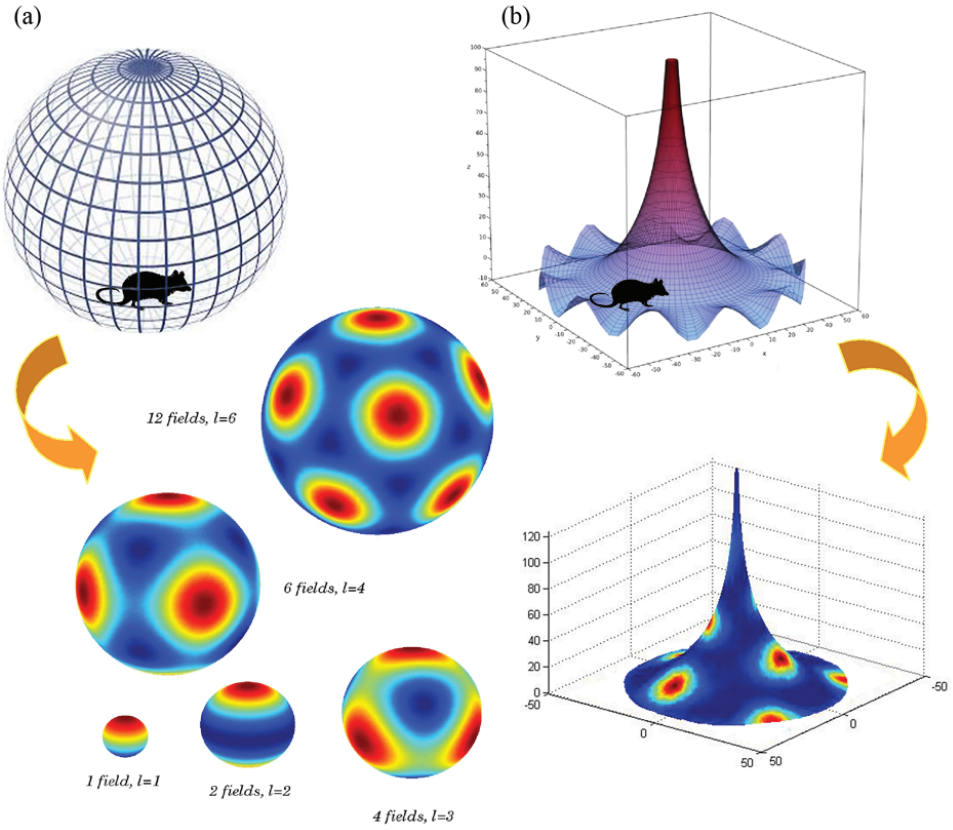


Figure 4: Grid patterns developed in non-flat surfaces. (a) Different grids predicted by the model when a rat is raised in a sphere (top), as the radius increases (keeping the grid spacing fixed). (b) The negatively curved counterpart of the sphere is the so-called pseudo-sphere. A rat nurtured in an hyperbolic cage (top) will develop patterns with a coordination number larger than 6.

We have also simulated rat development on a *pseudosphere*, an environment of constant but negative Gaussian curvature. The model predicts that the metric expressed by the conjunctive units should reflect the environment in which the animal develops and in fact we show that, if virtual

rats are raised on a pseudosphere, they form hyperbolic grids. For a given range of grid spacing relative to the radius of negative curvature of the hyperbolic surface, such grids appear as multi-peaked firing maps, in which each peak has seven neighbors instead of the Euclidean six, see Fig. 4.b for

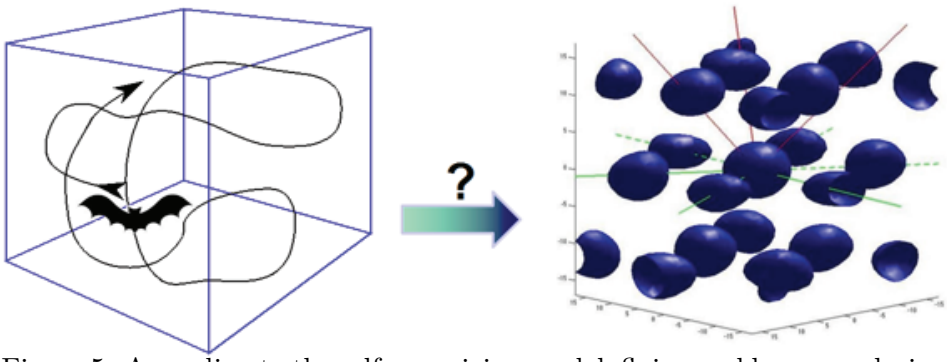


Figure 5: According to the self-organizing model, flying and hence exploring a 3D world for a sufficient time, a bat should generate volumetric grids, resembling the crystal-like organization of hard-sphere packings.

In the laboratory of Nachum Ulanovsky at the Weizmann Institute, on the other hand, putative grid units have been recorded in bats that fly in a large laboratory measuring approximately $5\times5\times3$ m^3. Again, we have simulated our conjunctive cell model to show that, given extensive exploration of a three dimensional volume, grid units can form with the approximate periodicity of a either a face-centered cubic crystal or an hexagonally compact packing. They appear again as the spontaneous product of a self-organizing process at the single unit level, driven solely by firing rate adaptation, while their alignment results from collateral interactions that also self-organize.

Acknowledgements

This work is supported by the EU FET project GRIDMAP, with contributions by Federico Stella, Francesca Troiani, Bailu Si and others in the project.

References

[1] J. O'Keefe and J. Dostrovsky, *Brain Res.* **34** (1971) 171.

[2] J. S. Taube, R. U. Muller and J. B. Ranck, *J. Neurosci.* **10** (1990) 420.

[3] T. Hafting, M. Fyhn, S. Molden, M. -B. Moser and E. I. Moser, *Nature* **436** (2005) 801.

[4] M. M. Yartsev, M. P. Witter and N. Ulanovsky, *Nature* **479** (2011) 103.

[5] A. Treves, O. Miglino and D. Parisi, *Psychobiology* **20** (1992) 1.

[6] M. Franzius, H. Sprekeler and L. Wiskott, *PLoS Comput. Biol.* **3(8)** (2007) e166.

[7] K. Zhang, *J. Neurosci.* **16** (1996) 2112.

[8] F. Sargolini, M. Fyhn, T. Hafting, B. L. McNaughton, M. P. Witter, M. -B. Moser and E. I. Moser, *Science* **312** (2006) 758.

[9] Freely distributed data from Moser lab can be found in http://www.ntnu.edu/kavli/research/grid-cell-data.

[10] C. N. Boccara, F. Sargolini, V. H. Thoresen, T. Solstad, M. P. Witter, E. I. Moser and M. -B. Moser, *Nat. Neurosci.* **13** (2010) 987.

[11] B. Si and A. Treves, *Hippocampus* **23** (2013) 1410.

[12] L. M. Giocomo, M. -B. Moser and E. I. Moser, *Neuron* **71** (2011) 589.

[13] E. A. Zilli, *Front. Neural Circuits* **6** (2012) 1.

[14] B. L. McNaughton, F. P. Battaglia, O. Jensen, E. I. Moser and M. -B. Moser, *Nat. Rev. Neurosci.* **7** (2006) 663.

[15] M. C. Fuhs and D. S. Touretzky, *J. Neurosci.* **26** (2006) 4266.

[16] Y. Burak and I. R. Fiete, *PLoS Comput. Biol.* **5** (2009) e1000291.

[17] Z. Navratilova, L. M. Giocomo, J. -M. Fellous, M. E. Hasselmo, B. L. McNaughton, *Hippocampus* **22** (2012) 772.

[18] N. Burgess, C. Barry and J. O'Keefe, *Hippocampus* **17** (2007) 801.

[19] L. M. Giocomo, E. A. Zilli, E. Fransén and M. E. Hasselmo, *Science* **315** (2007) 1719.

[20] N. Burgess, *Hippocampus* **18** (2008) 1157.

[21] M. E. Hasselmo, *Hippocampus* **18** (2008) 1213.

[22] A. Treves, E. Kropff and A. Biswas, *Soc. Neurosci. Abstract* **198** (2005) 11.

[23] E. Kropff and A. Treves, *Hippocampus* **18** (2008) 1256.

[24] B. Si, E. Kropff and A. Treves, *Biol. Cybern.* **106** (2012) 483.

[25] H. Mhatre, A. Gorchetchnikov and S. Grossberg, *Hippocampus* **22** (2012) 320.

[26] J. L. Kubie and A. A. Fenton, *Front. Neural Circuits* **6** (2012) 1.

[27] T. Bonnevie, B. Dunn, M. Fyhn, T. Hafting, D. Derdikman, J. L. Kubie, Y. Roudi, E. I. Moser and M. -B. Moser, *Nat. Neurosci.* **16** (2013) 309.

[28] R. F. Langston, J. A. Ainge, J. J. Couey, C. B. Canto, T. L. Bjerknes, M. P. Witter, E. I. Moser and M. -B. Moser, *Science* **328** (2010) 1576.

[29] T. J. Wills, F. Cacucci, N. Burgess and J. O'Keefe, *Science* **328** (2010) 1573.

[30] J. J. Couey, A. Witoelar, S. -J. Zhang, K. Zheng, J. Ye, B. Dunn, R. Czajkowski, M. -B. Moser, E. I. Moser, Y. Roudi and M. P. Witter, *Nat. Neurosci.* **16** (2013) 318.

[31] F. Stella, B. Si, E. Kropff and A. Treves, *J. Stat. Mech* **03** (2013) P03013.

Genome and Language – Two Scripts of Heredity

Edward N. Trifonov [a]
Genome Diversity Center, Institute of Evolution, University of Haifa, Israel

Abstract

One striking manifestation of life which is not easy to comprehend is invention of linear script, twice in evolution, first as genomic sequences, and then – as human language writings. Question to theory of informatics: is the simple linear script the best possible way to store and transmit the information? Evolution of both started, likely, from simple repetitions (TGTGTG…, GCCGCCGCC…, ma-ma-ma, da-da-da). Later the simple repetitions accumulated useful mutational changes, turning in more complex words already not recognizable as repeats. This scenario appears to be common for both genetic and language texts, which both seem to be, thus, of the same biological nature, carrying in different ways heritage of the human species.

[a] e-mail address: trifonov@research.haifa.ac.il

1 Introduction

Enigmatic and tantalizing is the fact that both genetic texts (nucleic acid and protein sequences) and human scripts are linear strings of symbols. Both carry information, each of its own kind, maintained in generations either by DNA replication (for genetic sequences), or by rewriting and reprinting (language texts). One can, of course, safely call it an analogy, however, the connection appears to be much deeper. Phenomenon of life as generator of genetic information has evolved on the basis of the sequences in the epoch of species formation, and on the same basis of linear script the life entered in its most advanced epoch - formation of human culture and languages.

The evolutionary nature of both genetic and language scripts suggests that, perhaps, there are some similarities in their origin and development. In particular, the nucleotide sequences are believed to evolve from simple tandem repeats [1], while human languages may have originated similarly, from primitive repetitions (baby babbling).

2 Origin and evolution of genomes

We start with the genome origin and evolution, described in several important details in a recent publication (ibid). The bottom line of this revealing study is that the genomes, especially eukaryotic ones, are full of repeat sequences – tandem repetitions of simple motifs, and dispersed repetitions of longer sequences. The repeats occupy more than 2/3 of human genome [2]. In the Table 1 the topmost repeats of human genome are presented [3]. It shows that simple repeats, like A_n, $(TG)_n$, dominate. Of the dispersed repeats dominant are the 15-mers from the Alu sequence, ~ 300 bases long, encountered in the genome every 3000 bases, in average. Such mass of the repeats does not seem to carry much of useful information. It appears rather that it is in the nature of the genomes to be constantly loaded by the repeats. And the question is why it is so. The clue is given by existence of so-called triplet expansion diseases, such that certain aggressively repeating triplets spontaneously increase their copy numbers (expand), causing neurodegenerative diseases and chromosome fragility. A speculation has been put forward

that such selfishly successful sequences have been always around, including the very first steps of emerging life [4]. Such successful sequences are likely to have been winners in the harsh competition in the earliest stages of life.

Table 1. Topmost 15-mers of human genome.

1	1 198 780	TTTTTTTTTTTTTTT	T_n
2	1 190 667	AAAAAAAAAAAAAAA	A_n
3	366 285	TGTGTGTGTGTGTGT	TG_n
4	362 623	ACACACACACACACA	AC_n
5	348 215	GTGTGTGTGTGTGTG	GT_n
6	344 421	CACACACACACACAC	CA_n
7	223 424	GCTGGGATTACAGGC	Alu
8	223 011	GCCTGTAATCCCAGC	Alu
9	222 894	TATATATATATATAT	TA_n
10	222 730	ATATATATATATATA	AT_n
11-67		fragments of Alu	
68	169 033	TTTTTTTTTTTTTTG	T_n
69-72		fragments of Alu	
73	167 889	CAAAAAAAAAAAAAA	A_n
74	167 361	CTAAAAATACAAAAA	Alu
75	150 349	CTTTTTTTTTTTTTT	T_n
76	149 748	AAAAAAAAAAAAAAG	A_n
77-82		fragments of Alu	

The straightforward hypothetical scenario for the gene and genome origin would be spontaneous appearance of the repeats in the primitive early genomes, and their subsequent mutational changes towards higher sequence complexity while freshly generated repeats would continue to invade the genomes. As the very first duplex gene the aggressive sequences GCC_n and complementary GGC_n have been suggested (ibid). Analysis of very large amount of triplets of mRNA demonstrated that the aggressive triplet expansion sequences are massively present in the mRNA, in diversified but well recognizable form [1]. In other words, the memory about original repeat expansion events still survives in the mRNA sequences. This is also true for sequences not coding for proteins [5].

3 Remarks on origin and evolution of language

Thus, genomes appear to emerge and evolve by spontaneous repeat expansions and concomitant diversification of the repeats to higher sequence complexity. Origin and evolution of the second script of life – language – may have had the same basic characteristics. In particular, one would expect the language to start with simple repetitions as well, and continue into further diversification. Evolutionary intuition suggests that "Language owes its origin to the imitation and modification,…, of various natural sounds, the voices of other animals, and man`s own instinctive cries" [6]. Following this lead we now turn to the ontogenetic theory of the origin of the human language, developing it from the "instinctive cries" of babies, with their simple speech apparatus, perhaps, as simple as in ancestral hominids.

In second half of 19th century an exciting idea was put forward, by Ernst Haeckel [7], that the consecutive stages of development of animal embryos were very close to respective evolutionary stages. For example, human embryo undergoes transformations from what at some point apparently corresponds to fish, and later such features as vertebrae, four extremities and others sequentially appear. This, indeed, parallels fairly well the evolutionary stages of mammals: cartilagenous fish – bony fish – four-legged animals. One can imagine that some later evolutionary advances also in some degree obey the ontogeny/phylogeny trend. In particular, the speech apparatus of the embryo and of infant may well be

simpler at the beginning, different from the adult structures in terms of detailed anatomy and basic muscles involved. Perhaps, the same was true for adult ancestral hominids. It is very tempting to assume that the first "speaking" hominid ancestors, with only those speech muscles, which were formed at their stage of evolution, managed to pronounce primarily those consonants which were easier to pronounce. These earliest speech muscles, presumably, are also those which are formed first in the human ontogenesis. In other words, the speech apparatus of a child, in its development from postnatal to later stages, basically, repeats all steps of evolution from first primitive speakers to modern human, starting with the very first simple repeat vocabulary, presumably, common for all babies of speaking hominids of all times.

4-7 month old babies can pronounce the consonants p, b, t, m, d, n, k, g, s, h, w, j (baby consonants) , while f, v, th, sh, ch, l, r are problematic at this stage of development of the speech apparatus [8]. Importantly, the pronounceability of various consonants by babies is the same irrespective of the language environment and ethnicity of the babies (ibid), thus, apparently being function of the available arsenal of speech muscles, which is the same in babies of the same age, for all ethnicities.

Every mother is fascinated by the sounds uttered by the babies in their first months, initially in no connection with outer world, no conscious communication. European mother, of course, would enthusiastically respond to spontaneous "ma-ma-ma", thus, establishing and further consolidating the first liaison of baby words with reality. Georgian mother would react the same way to "da-da-da" ("dada" is mother in Georgian), while Swahili speaking mother would respond to "ba-ba-ba". The repeating syllables, "words" with the baby consonants in practically all languages have also baby meanings, describing their closest environment: mother, father, grandmother, grandfather, food, milk, breast, feces etc. In the Table I the repeat words-syllables containing the easily pronounceable consonants are presented, in various languages. The repeating consonants alternating with vowels are very similar to the vocabulary of earliest sequence repeats in genomes (GCC-GCC-GCC, GGC-GGC-GGC,… , TG-TG-TG, A-A-A-A,… and alike).

Table 2. Examples of easily pronounceable baby words in various languages. (From various sources, including interviews with experienced mothers).

Baba	baby (Arabic), father (Bengali, Mandarin, Swahili), grandmother (Russian)
Mama	mother (Arabic, Mandarin, Russian, Swahili), father (Georgian), food (Japanese), breast milk (Malayan)
Kaka	feces (in many languages)
Nana	food (Arabic), mother (Fijian), father (Telugu)
Papa	father (European), grandfather (Georgian), mother (Japanese), baby food (Malayan)
Sisi	bird (Arabic), breast (Russian)
Tutu	dog (Arabic), locomotive (Russian)

From the simple repeats and their mutations the evolution of language, likely, entered the next pronounceability stage – alternation of mixed consonants and vowels (like "capability"). Similar alternations are characteristic also for protein sequences where the alternating polar and non-polar amino acid residues correspond to amphipathic alpha-helices of proteins [9].

Thus, the early language has been, likely, gradually progressing from the repeats of easiest pronounceability to more diverse repeats (also involving difficult consonants), to heterogeneous alternations of consonants and vowels, and, finally, to a whole spectrum of words. And, again, this parallels the diversification of nucleotide and protein sequences from simple repeats to complex sequences. In this case the functional utility of the appearing new mutated forms serves as the pronounceability in language. This gradient from words easily tackled to more complicated words, from simple to complex, is the most natural trend in evolution, common for both genetic and language scripts. "...from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved" [10].

4 Language as biological phenomenon

We fully realize that the biological connection of the languages is highly controversial issue. The connection, however, is very obvious, as it follows from what is described above. There are plenty of characteristics shared by the two scripts of different(?) nature, such as frequency vocabularies and rules of "pronounceability". There are, of course, some notable differences as well. One of them is overlapping character of the multiple codes in genomes [11], while such overlapping in languages is only rare, exotic phenomenon (like acrostichs). Another important difference is strict rule of almost ideal identity in replication, while the identity in writings is, essentially, kept only for consecutive editions of canonical and authored texts. Diversity in human writings is, virtually, unlimited, while only certain degree of diversity is allowed in evolution of organisms. Questions to social sciences: May one consider language and the whole corpus of writings not just a cultural heredity but also as subject of biological (social) evolution? Is there pressure of natural selection acting on ideas and canonical texts (basic human knowledge, religious writings, theories)? Will they eventually crystallize in frozen invariants? The tempting, perhaps, non-orthodox thought is: both types of texts are products and subjects of evolution, inseparable manifestations of life, both belonging to the domain of biology, both representing heritage of *H. sapiens*, genetic and cultural. And both components of the heritage are crucially important for identity and survival of the human beings.

References

[1] Z. M. Frenkel and E. N. Trifonov, J. Biomol. Str. Dyn. 30 (2012) 201.

[2] A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer and D. D. Pollock, PLoS Genet. 7 (2011) e1002384.

[3] A. E. Rapoport and E. N. Trifonov, J. Biomol. Str. Dyn. 31 (2013) 1324.

[4] E. N. Trifonov and T. Bettecken, Gene 205, (1997) 1.

[5] E. N. Trifonov, Z. Volkovich and Z. M. Frenkel, Annals NY Acad. Sci. 1267 (2012) 35.

[6] C. Darwin, The descent of man, and selection in relation to sex. London, Murray, 1871, p. 56.

[7] E. Haeckel, The Wonders of Life: A Popular Study of Biological Philosophy, London, Watts & Co., 1904.

[8] W. O`Grady, J. Archibald, M. Aronoff and J. Rees-Miller, Contemporary Linguistics: An Introduction, Bedford/St. Martin's, 2000.

[9] M. Zemkova, E. N. Trifonov and D. Zahradnik, J. Biomol. Str. Dyn. DOI:10.1080/07391102.2013.809317 (2013)

[10] C. Darwin, On the origin of species, London. Murray, 1859.

[11] E. N. Trifonov, Bull. Math. Biol. 51 (1989) 417.

Simple Physics and Bioinformatics of Nucleosome Positioning

Edward N. Trifonov [a]
Genome Diversity Center, Institute of Evolution, University of Haifa, Israel

Abstract

The problem of sequence-specific nucleosome positioning exists since 1980, when first evidence has been obtained that chromatin DNA has a hidden 10-11 base periodicity, apparently associated with DNA bending in the nucleosome. Since chromatin community has no background in weak signal processing, the field of nucleosome positioning suffered three decades of mistrust, confusion and misconceptions. One especially damaging wrong idea was mirror symmetry of the nucleosome positioning sequence pattern massively mistaken for complementary dyad symmetry, although the physically correct picture has been described many times since 1980, as well as the sequence patterns consistent with the physics of DNA deformation. Recent discovery of strong nucleosome DNA sequences with clearly visible rather than hidden sequence periodicity, hopefully, puts an end to the misunderstanding. Deformation of DNA in the nucleosome is largely guided by the RR/YY dinucleotide stacks. The purine residues are harder to unstack,

[a] e-mail address: trifonov@research.haifa.ac.il

therefore they are placed towards DNA-histone interface. This makes the alternation R5Y5 an ideal sequence for DNA bending in the nucleosome and for sequence-dependent nucleosome positioning. The whole length nucleosome DNA positioning pattern oscillates with the period 10.4 bases. It allows accurate mapping of the nucleosomes along the sequences with single-base resolution, so that orientation and accessibility of sequence elements of interest on the surface of the nucleosome are also specified.

1 Nucleosomes

The first observation introducing the notion of chromatin structural unit was the one by Hewish and Burgoyne [1] who discovered that the endonuclease digests of chromatin generate a ladder of discrete DNA fragments, apparently monomers, dimers and higher oligomers. The minimal monomer size of 146 base pairs universal for all eukaryotes became a primary feature in an operational definition of the nucleosome. Subsequent physical chemistry studies added histone octamers as protein component of the nucleosome [2]. In 70's it also became clear that the nucleosome is not indifferent to DNA sequence. Moreover, at the preferential regions of the binding of histone octamer to DNA there are several alternative positions for the binding, separated by 10 or 11 bases (e.g., [3]). The 10-11 base periodicity of the nucleosome DNA sequence apparently associated with DNA bending in the nucleosome [4] is second major operational definition of the nucleosome. During decades following this discovery the chromatin community agonized through uncertainties of the phenomenon, largely because the periodic signal is very weak, and special signal processing tools have to be used to detect it and to characterize sequence-wise, i.e. to establish the periodically repeating sequence pattern(s). Many sequence motifs have been suggested, often in disaccord with one another [5]. This is understandable, as only minute fraction of the chromatin research community has some knowledge of signal processing, especially when the signal is as weak as in nucleosome DNA. Indeed, until very lately none of the known nucleosome DNA sequences displayed any obvious pattern which could be recognized as periodic. Moreover, since strongly periodic sequences would, perhaps, make strong nucleosomes, one would expect that these should be massively avoided, not to interfere with replication and transcription of DNA. Thus, the signal has to be of moderate strength at the best, just to indicate the optimal nucleosome positions amongst many possible comparable alternatives. The selective affinity of histone octamers to the specific (best) positions along the nucleosome DNA sequence is spectacularly demonstrated by crystallization of the nucleosomes reconstructed on specific sequences which resulted in identification of bases located at the very dyad axis of

the nucleosome, for every sequence taken for such experiment (reviewed in [6]). And, again, no sequence periodicity is visible in the crystallized sequences, though they are accurately placed in the nucleosomes, guided by the (hidden) sequence periodicity [7].

The field of nucleosome positioning suffers already more than three decades from mistrust, confusion and misconceptions. One especially misleading idea was mirror symmetry of the nucleosome positioning sequence pattern, mistaken for complementary dyad symmetry. The work of Satchwell et al. [8] unintentionally contributed to the misunderstanding. In it, the occurrences of various dinucleotides along the nucleosome DNA were analyzed as "functions of their position in the core DNA *molecule*". In other words, the occurrences have been registered along the DNA *duplex*, i. e. in both complementary strands simultaneously. If the dinucleotide AA is preferentially found at position x in one strand (say, Watson-strand, AA_{Watson}), the complementary TT_{Crick} of the other strand is there as well, at the same distance x from the dyad axis. That is, $AA_{Watson}(x) = TT_{Crick}(x)$. This trivial equality has been understood by most of researchers as *AA and TT are interchangeable,* as if their physical properties, when they are placed in the same position *of the same strand*, are identical. This is, certainly, not the case if, say, deformational potential (stacking of neighboring A and neighboring T) is considered. The plots in [8] are deceptively mirror-symmetrical, not reflecting the true symmetry characteristic of the DNA duplex - complementary, dyad symmetry. That is, if there is preferential position x for AA in the nucleosome DNA sequence, the corresponding preferential position for TT is –x. AA(x) = TT(-x), but not AA(x) = TT(x). Peaks for AA in the distribution along the averaged nucleosome DNA sequence are complementarily symmetrical to peaks for TT. The 1986 AA=TT work [8] is still frequently referred to by massively confused chromatin community, although the physically correct picture has been described many times since 1980, as well as the sequence patterns consistent with the physics of DNA deformation and dyad symmetry [9].

The main physical factor responsible for exact DNA positioning in the nucleosome, i.e., bending in one specific direction is stacking interactions between neighboring bases and base pairs. There are four types of base-pair stacks: Purine-Purine (RR) complementary to Pyrimidine-Pyrimidine (YY) and *vice versa*, i.e., RR*YY and YY*RR stacks, Purine-Pyrimidine (RY) complementary to Purine-Pyrimidine as

well – RY*RY stack, and YR*YR stack, represented in the sequence as RR, YY, RY and YR dinucleotides. The first two types are very asymmetric in terms of base-to-base stacking [10]. The purine residues (A and G) following one another in the sequence are harder to unstack, compared to neighboring pyrimidine residues (C and T). When DNA is deformed on the surface of histone octamer the purines should be placed towards DNA-histone interface, while pyrimidines should be oriented away from the surface, to minimize the energy required for the deformation. This is illustrated in the Fig. 1, where the purines (dark gray) are, essentially, below the DNA axis, while pyrimidines (light gray) are above. This makes the alternation R_5Y_5 an ideal sequence for DNA bending in the nucleosome. In addition, the YR dinucleotides are known to be preferentially located in minor grooves of DNA facing the histone octamer [11], as seen in the scheme.

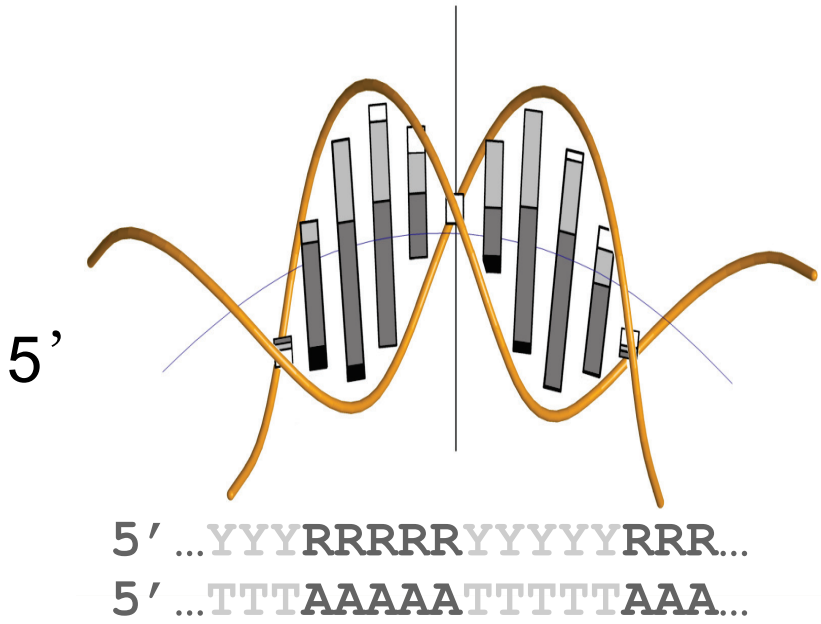


Figure 1. Schematic view of one helical repeat of DNA on the surface of the histone octamer. Purines (R) – dark grey, Pyrimidines (Y) – light grey.

A non-trivial issue in the nucleosome structure is the mean value of the DNA period in the nucleosome (reviewed in [12]). The period is non-integer, which is *per se* a stumbling block for chromatin research community. The earliest 1979 estimates of the period, 10.3-10.5 base pairs per one turn of the duplex [13] are very close to the most accurate later estimates, all converging to 10.36-10.40 bp/turn [12]. The correct value of the period is very important for derivation of the nucleosome DNA consensus sequence for accurate sequence-directed nucleosome mapping.

Most previous techniques used for the extraction of the nucleosome DNA periodical pattern ended with the dominating $(RRRRRYYYYY)_n$ sequence [14] in accordance with the simple physics above, with 10 or 11 bases within parentheses, to satisfy average 10.4 base periodicity. The pattern and its matrix version [15] allow accurate mapping of the nucleosomes along the sequences with single-base resolution, so that orientation and accessibility of sequence elements of interest on the surface of the nucleosome are also specified [16].

2 Strong nucleosomes

The concept and the discovery of the strong nucleosomes, SNs [12] breathed a new life in the problem of the nucleosome sequence pattern. The honor of the very first demonstration that DNA sequence periodicity imparts to nucleosomes higher stability, makes them strong, belongs to Lowary and Widom [17] who also produced the first comparatively strong nucleosome DNA, clone 601, for further use in biologically motivated experiments. The idea of the experiment was both simple and daring: find within large ensemble of random (unnatural) DNA sequences those which after reconstruction with histone octamers make nucleosomes resistant to dissociation. The sequences revealed a high proportion of TA dinucleotides following one after another at a distance of 10 bases. We repeated the experiments of Lowary and Widom computationally, but on natural sequences, by looking for exceptionally periodic, perhaps, visibly periodic sequence segments in eukaryotic genomes. Such sequences have been found indeed, and, naturally, named Strong Nucleosomes, SNs [12]. Never before the nucleosome positioning signal have been seen with such degree of obviousness.

In the Fig. 2 examples of the SNs from small plant *Arabidopsis thaliana* are presented, with visible periodical appearance of TTTTAAAA sequence or its minimally modified versions, with the period 10-11 bases. More detailed analysis, with application of dynamic programming to over 500 SN sequences resulted in the consensus ATTTTTAAAAAT (TA-central) or TAAAAATTTTTA (AT-central). The procedure of finding the most periodic sequences was not geared to any specific dinucleotides. Contribution of each of 16 dinucleotides to the oscillation with the period 10.4 bases was determined, and total sum of the amplitudes scored for every fragment along sequence of interest, thus, resulting in the detection of the SN sequences.

Figure 2. Small sample of aligned strong nucleosome DNA sequences of *A. thaliana*. Alternating with period ~10.4 bases runs of A (red) and T (blue) are clearly visible.

As it would follow from the previous section, the natural sequences should avoid any DNA segments forming strong nucleosomes. And, indeed, the SNs are very rare species – their concentration in three different genomes analyzed (*A. thaliana, C. elegans, H. sapiens*) is of the order of 1 per megabase, i.e., about 1 per 5000 ordinary nucleosomes. This unique population of SNs should have some very special function.

And, indeed, they are preferentially found in centromeric regions of chromosomes [18]. Centromeres are specific sites in chromosomes, where the homologous pairs of chromosomes contact each other.

The SNs have been extracted from several sequence ensembles, and the corresponding sequence consensuses of the repeats have been generated, ending with $(AGAGGCCTCT)_n$ for sequences of strong nucleosomes in experiments of Lowary and Widom, $(AAAAATTTTT)_n$ for A. thaliana, C. elegans and H. sapiens (total), and $(GGGGGCCCCC)_n$ for G+C rich isochores H3 of H. sapiens. Remarkably, all three consensuses above are, actually, $(RRRRRYYYYY)_n$, in full agreement with earlier studies involving ordinary nucleosomes [14]. In other words, the hidden pattern in ordinary nucleosomes is the same as visible periodical pattern in SNs, so that the SNs, with much stronger periodic signal, can be used now for further detailed studies of the 4-letter alphabet variants of the RRRRRYYYYY nucleosome positioning motifs.

References:

[1] D.R.Hewish and L.A. Burgoyne, Bioch. Bioph. Res. Comm. 52 (1973) 504.

[2] K. E. van Holde, C. G. Sahasrabuddhe and B. R. Shaw, Nucl. Acids Res. 1 (1974) 1579; R. D. Kornberg, Science 184 (1974) 868.

[3] B. A. J. Ponder and L. V. Crawford, Cell 11 (1977) 35.

[4] E. N. Trifonov and J. L. Sussman, Proc. Natl. Acad. Sci. USA 77 (1980) 3816.

[5] E. N. Trifonov, J. Biomol. Str. Dyn. 27 (2010) 741.

[6] T. J. Richmond and C. A. Devey, Nature 423 (2003) 145.

[7] I. Gabdank. D. Barash and E. N. Trifonov, J. Biomol. Str. Dyn. 28 (2010) 107.

[8] S. C. Satchwell, H. R. Drew and A. A. Travers, J. Mol. Biol. 191 (1986) 659.

[9] G. Mengeritskly and E. N. Trifonov, Nucl. Acids Res. 11 (1983) 3833; E. N. Trifonov, J. Theor. Biol. 263 (2010) 337; E. N. Trifonov, Phys. Life Rev. 8 (2011) 39.

[10] S. Arnott, S. D. Dover and A. J. Wonacott, Acta Cryst. B25 (1969) 2192; E. N. Trifonov, CRC Crit. Rev. Bioch. 19 (1985) 89.

[11] F. Cui and V. B. Zhurkin, J. Biomol. Str. Dyn. 27 (2010) 821; E. Y. D. Chua, D. Vasudevan, G. E. Davey, B. Wu and C. A. Davey, Nucl. Acids Res. 40 (2012) 6338.

[12] B. Salih, V. Tripathi and E. N. Trifonov, J. Biomol. Str. Dyn. (2013) DOI: 10.1080/07391102.2013.855143.

[13] A. Prunell, R. Kornberg, L. Lutter, A. Klug, M. Levitt and F.H.C. Crick, Science 204 (1979) 855; E. N. Trifonov and T. Bettecken, Biochemistry 18 (1979) 454.

[14] G. Mengeritskly and E. N. Trifonov, Nucl. Acids Res. 11 (1983) 3833; F. Salih, B. Salih and E. N. Trifonov, J. Biomol. Str. Dyn. 26 (2008) 273; A. E. Rapoport, Z. M. Frenkel and E. N. Trifonov, J Biomol Struct Dyn 28 (2011) 567; I. Gabdank, D. Barash and E. N. Trifonov, J. Biomol. Struct. Dyn. 26 (2009) 403.

[15] I. Gabdank, D. Barash and E. N. Trifonov, J. Biomol. Struct. Dyn. 26 (2009) 403; J. Biomol. Str. Dyn. 28 (2010) 107.

[16] J. Hapala and E. N. Trifonov, Gene 489 (2011) 6; Gene 527 (2013) 339.

[17] P. T. Lowary and J. Widom, J. Mol. Biol. 276 (1998) 19.

[18] B. Salih and E. N. Trifonov, J. Biomol. Struct. Dyn. (2013) DOI: 10.1080/07391102.2013.860624; J. Biomol. Struct. Dyn. (2014) DOI:10.1080/07391102.2013.879263.

The Author Index