Proceedings

Belgrade BioInformatics Conference 2016

20-24 June 2016, Belgrade, Serbia

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Nenad Mitić, editor

Belgrade BioInformatics Conference 2016

Proceedings

Belgrade, June 20th-24th

The conference is organized by the Bioinformatics Research Group, University of Belgrade - Faculty of Mathematics (http://bioinfo.matf.bg.ac.rs).

Coorganizers of the conference are: Faculty of Agriculture, Faculty of Biology, Faculty of Chemistry, Faculty of Physical Chemistry, Institute for Biological Research "Siniša Stanković", Institute for General and Physical Chemistry, Institute for Medical Research, Institute of Molecular Genetics and Genetic Engineering, Vinča Institute of Nuclear Sciences, Mathematical Institute of SASA, Belgrade, and COST - European Cooperation in Science and Technology

The conference is financially supported by

- Ministry of Education, Science and Technological Development of Republic of Serbia
- Central European Initiative (CEI)
- Telekom Srbija
- SevenBridges Genomic
- RNIDS - Register of National Internet Domain Names of Serbia
- Genomix4Life

Publication of this Proceedings is financed by the Ministry of Education, Science and Technological Development of Republic of Serbia

Publisher: Faculty of Mathematics, University of Belgrade
Printed in Serbia, by DonatGraf, Belgrade

Serbian National Library Cataloguing in Publication Data
Faculty of Mathematics, Belgrade
Proceedings: Belgrade BioInformatics Conference 2016, 20-24 June 2016.– Proceedings
Nenad Mitić, editor. XII+181 pages, 24cm.

Publication year: 2017.

ISBN: 978-86-7589-124-6

Number of copies printed: 100

International Advisory Committee

Vladik Avetisov	The Semenov Institute of Chemical Physica, RAS Moscow, Russia
Vladimir Brusić	School of Medicine and Bioinformatics Center, Nazarbayev University, Kazakhstan and Department of Computer Science, Metropolitan College, Boston University, USA
Michele Caselle	Department of Physics, Torino University, Torino, Italy
Radu Constantinescu	Department of Physics, University of Craiova, Craiova, Romania
Oxana Galzitskaya	Group of bioinformatics, Institute of Protein Research of the RAS, Russia
Madhavi Ganapathiraju	Department of Biomedical Informatics, University of Pittsburgh, USA
Mikhail Gelfand	A.A. Kharkevich Institute for Information Transmission Problems, RAS, Faculty of Bioengineering and Bioinformatics, M.V. Lomonosov Moscow State University, Moscow, Russia
Ernst Walter Knapp	Fachbereich Biologie, Chemie, Pharmazie/Institute of Chemistry and Biochemistry, Freie Universitt Berlin, Germany
Sergey Kozyrev	Steklov Mathematical Institute, Moscow, Russia
Zoran Obradović	Center for Data Analytics and Biomedical Informatics, Temple University, USA
Yuriy L. Orlov	Institute of Cytology and Genetics SB RAS, Novosibirsk State University, Russia
George Patrinos	Department of Pharmacy, University of Patras, Greece
Nataša Pržulj	Department of Computing , Imperial College London, UK
Paul Sorba	Laboratory of Theoretical Physics and CNRS, Annecy, France
Bosiljka Tadić	Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, Slovenia
Peter Tompa	VIB Structural Biology Research Center, Flanders Institute for Biotechnology (VIB), Belgium
Silvio Tosatto	Department of Biomedical Sciences, University of Padova, Italy
Edward Trifonov	Weizmann Institute of Science, University of Haifa, Haifa, Israel
Matthias Ullmann	Structural Biology/Bioinformatics Universitt Bayreuth, Germany
Bane Vasić	The University of Arizona, Department of Electrical and Computer Engineering, Bios Institute for Collaborative Bioresearch, USA
Sergey Volkov	Bogolyubov Institute for Theoretical Physics, Kiev, Ukraine
Ioannis Xenarios	SIB Swiss Institute of Bioinformatics, Switzerland

International Programme Committee

Miloš Beljanski	Institute for General and Physical Chemistry, University of Belgrade, Serbia
Erik Bongcam-Rudloff	Division of Molecular Genetics, Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Sweden
Antonio Cappuccio	Immunity and Cancer, Institut Curie, France
Oliviero Carugo	Faculty of Science, University of Pavia, Italy
Boris Delibašić	Faculty of Organizational Sciences, University of Belgrade, Serbia
Zsuzsanna Dosztanyi	Department of Biochemistry Eötvös Loránd University, Budapest, Hungary
Branko Dragovich	Institute of Physics, Mathematical Institute SANU, Belgrade, Serbia
Marko Djordjević	Faculty of Biology, University of Belgrade, Serbia
Olgica Djurković-Djaković	Institute for Medical Research, University of Belgrade, Serbia
Lajos Kalmar	Department of Veterinary Medicine, Cambridge Veterinary School, Cambridge, UK
Eija Korpelainen	CSC IT Center for Science, Finland
Ilija Lalović	Faculty of Natural Sciences and Mathematics, Banja Luka, Bosnia and Herzegovina
Nenad Mitić	Faculty of Mathematics, University of Belgrade, Serbia
Mihajlo Mudrinić	Vinča Institute of Nuclear Sciences, University of Belgrade, Serbia
Zoran Ognjanović	Mathematical Institute SANU, Serbia
Gordana Pavlović-Lažetić	Faculty of Mathematics, University of Belgrade, Serbia
Marco Punta	Pierre and Marie Curie University, France
Predrag Radivojac	Department of Computer Science and Informatics, Indiana University, USA
Ana Simonović	Institute for Biological Research Siniša Stanković, Belgrade, Serbia
Jerzy Tiuryn	Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
Andrew Torda	Center for Bioinformatics, University of Hamburg, Germany
Alessandro Treves	SISSA-Cognitive Neuroscience, Trieste, Italy
Nevena Veljković	Institute for Nuclear Sciences VINCA, University of Belgrade, Serbia
Igor V. Volovich	Department of Mathematical Physics, Steklov Mathematical Institute, RAS, Moscow, Russia
Snežana Zarić	Faculty of Chemistry, University of Belgrade, Serbia

Local Organizing Committee

Bojana Banović	Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Miloš Beljanski	Institute for General and Physical Chemistry, University of Belgrade, Serbia
Branko Dragovich	Co-Chair, Institute of Physics, Mathematical Institute SANU, Belgrade, Serbia
Marko Djordjević	Faculty of Biology, University of Belgrade, Serbia
Olgica Djurković-Djaković	Institute for Medical Research, University of Belgrade, Serbia
Jelana Guzina	Faculty of Biology, University of Belgrade, Serbia
Jovana Kovačević	Faculty of Mathematics, University of Belgrade, Serbia
Saša Malkov	Faculty of Mathematics, University of Belgrade, Serbia
Mirjana Maljković	Faculty of Mathematics, University of Belgrade, Serbia
Vesna Medaković	Faculty of Chemistry, University of Belgrade, Serbia
Nenad Mitić	Co-Chair, Faculty of Mathematics, University of Belgrade, Serbia
Ivana Morić	Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Mihajlo Mudrinić	Vinča Institute of Nuclear Sciences, University of Belgrade, Serbia
Vesna Pajić	Faculty of Agriculture, University of Belgrade, Serbia
Mirjana Pavlović	Institute for General and Physical Chemistry, University of Belgrade, Serbia
Gordana Pavlović-Lažetić	Co-Chair, Faculty of Mathematics, University of Belgrade, Serbia
Jelena Samardzić	Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Ana Simonović	Institute for Biological Research Siniša Stanković, Belgrade, Serbia
Miomir Stanković	Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia
Biljana Stojanović	Faculty of Mathematics, University of Belgrade, Serbia
Aleksandra Uzelac	Institute for Medical Research, University of Belgrade, Serbia

Preface

This book contains 18 papers related to talks or posters presented at the Belgrade Bioinformatics Conference 2016 (BELBI 2016), held 20 - 24 June 2016 in Belgrade, Serbia. We are grateful to all authors for writing their contributions to these proceedings. Articles presented here should be useful not only to participants of this conference but also to PhD students and other researches in bioinformatics and related topics.

This international conference grew out of the communities of previous conferences held in Belgrade, Data Mining in Bioinformatics (DMBI, 2012) and the Theoretical Approaches to Bioinformation Systems (TABIS, 2010, 2013). It was organized by the Bioinformatics group from the University of Belgrade, Faculty of Mathematics, in cooperation with several other institutions from the University of Belgrade (Faculty of Agriculture, Faculty of Biology, Faculty of Chemistry, Faculty of Physical Chemistry, Institute for Biological Research "Siniša Stanković", Institute for General and Physical Chemistry, Institute for Medical Research, Institute of Molecular Genetics and Genetic Engineering, Vinča Institute of Nuclear Sciences), Mathematical Institute of the Serbian Academy of Science and Arts and COST (European Cooperation in Science and Technology) Action BM1405.

The main purpose of the BelBI 2016 conference was illumination of different aspects of bioinformation systems, from theoretical approaches to modeling different phenomena in life sciences, to information technologies necessary for analysis and understanding huge amount of data generated, to application of computer science and informatics in the domain of precision medicine, finding new remedies against debilitating diseases and drug development. This is a big interdisciplinary and interrelated field of research.

The conference focused on three main research fields:

1. Theoretical Approaches to BioInformation Systems (TABIS).
2. Bioinformatics and Data Mining for OMICs Data.
3. Biomedical Informatics.

The conference program contained keynote lectures, invited talks, selected oral and poster presentations.

We thank all members of the International Advisory, Program and Organizing Committees for their help to have this event successful. We also thank all speakers for their high level, interesting and valuable talks, and other participants for their active attendance. We express our gratitude to the sponsors: Ministry of Education, Science and Technological Development of the Republic of Serbia; Central European Initiative(CEI); Telekom Srbija; SevenBridges Genomic; RNIDS - Register of National Internet Domain Names of Serbia; and Genomix4Life.

The overall number of participants was 123 and they came from 24 countries. We hope that all attending this conference will remember it as a useful and

pleasant event, and will wish to participate again in the future.

Some more information on BELBI 2016 can be found at the conference website http://www.belbi2016.matf.bg.ac.rs/. The second conference in this series of conferences, BELBI 2018, will be organized in the similar way and will be held 18 - 22 June 2018 in Belgrade, see http://belbi.bg.ac.rs/.

Belgrade, December 2017

Editors:
Branko Dragovich
Gordana Pavlovic-Lazetić
Nenad Mitić

BelBI2016, Belgrade, June 2016.

Table of Contents

Protein dynamics and protein functioning 1
Vladik Avetisov and Ekaterina Borshcheva

Radiation Induced Dysfunctions in the Working Memory Performance Studied by Neural Network Modeling 18
A.N. Bugay, G.F. Aru, E.B. Dushanov, and A.Yu. Parkhomenko

On Similarity related to the Genetic Code 29
Branko Dragovich and Nataša Ž. Mišić

Machine learning-based approach to help diagnosing Alzheimer's disease through spontaneous speech analysis 38
Jelena Graovac, Jovana Kovačević, and Gordana Pavlović Lažetić

Improving 1NN strategy for classification of some prokaryotic organisms . 43
M. Grbić, A. Kartelj, D. Matić, and V. Filipović

T-cell epitope prediction, the influence of amino acids physicochemical propeterties and frequencies on identifying MHC binding ligands 55
Davorka R. Jandrlić, Nenad S. Mitić, and Mirjana D. Pavlović

Networks of Interaction in Moving Animal Groups and Collective Changes of Direction 64
Asja Jelić

Filtering of repeat sequences in genomes 73
Ana Jelović, Miloš Beljanski, and Nenad Mitić

Could integrative bioinformatic approach predict the circulating miRs that have significant role in pancreatic tissue in type 2 diabetes? 82
Ivan Jovanović, Maja Živković, Jasmina Jovanović, Tamara Djurić, and Aleksandra Stanković

A biologically-inspired model of visual word recognition 88
Yair Lakretz, Naama Friedmann, and Alessandro Treves

A Quantum Approach to the DNA Functioning 93
A. Nicolaidis

Mining PMMoV genotype-pathotype association rules from public databases 102
Vesna Pajić, Bojana Banović, Miloš Beljanski and Dragana Dudić

Intermittency-driven complexity in the brain: towards a general-purpose event detection algorithm 108
Paolo Paradisi, Marco Righi, Umberto Barcaro, Ovidio Salvetti, Alessandra Virgillito, Maria Chiara Carboncini, and Laura Sebastiani

A Mathematical description of the Genetic Code: Symmetry and Minimum Principle 119
A. Sciarrino and P.Sorba

Algebraic topology of multi-brain graphs: Methods to study the social impact and other factors onto functional brain connections 134
Bosiljka Tadić and Miroslav Andjelković

Gene expression in schizophrenia patients and non-schizophrenic individuals infected with *Toxoplasma gondii* 142
Aleksandra Uzelac, Tijana Štajner, Miloš Busarčević, Ana Munjiza, Milutin Kostić, Čedo Miljević, Dušica Lečić-Toševski, Nenad Mitić, Saša Malkov, and Olgica Djurković-Djaković

Viral: Real-world competing process simulations on multiplex networks . . 151
Petar Veličković, Andrej Ivašković, Stella Lau, and Miloš Stanojević

DNA deformations as a tool for the genetic information implementation . . 159
Sergey N. Volkov

Author Index 171

Sponzors 175

Protein dynamics and protein functioning

Vladik Avetisov[1,2] and Ekaterina Borshcheva[1,2]

[1] The Semenov Institute of Chemical Physics of the Russian Academy of Sciences, Kosygina 4, 119991 Moscow, Russia
[2] National Research University Higher School of Economics, Myasnitskaya 20, 101000 Moscow, Russia
vladik.avetisov@gmail.com; katbr@yandex.ru

Abstract. In this paper, we present the studies of an ultrametric mathematical model for protein operation and give them physical interpretations that extend the conventional view of enzymatic activity regulation. The model is based on a representation of a multidimensional rugged energy landscapes by a hierarchy of nested basins of local minima and an approximation of protein dynamics with an ultrametric random walk. In contrast to an ordinary random walk, the ultrametric random walk is more suitable for describing of multiscale conformational dynamics and it is consistent with the kinetic features of ligand binding. Using our ultrametric model we show different ways to regulate enzymatic activity.

Keywords: ultrametric models, protein dynamics, protein functioning, enzymatic activity, regulation

1. Introduction

Proteins are one of the key biopolymers in a cell. They are often regarded as molecular machines that carry out various operations at the molecular level, e.g. specific binding, charge transfer, formation or breaking of a chemical bond, *etc*. It has long been understood that the outstanding functional ability of proteins is due not to only their specific folding, but also of their specific conformational mobility (for earlier views see, for example, [1]). Among globular polymers, proteins are distinguished by having a very wide range of intra-structural movements typical for soft matter, from one side, and for solid matter, from the other side. When we speak about picoseconds timescales we have in mind the motility of relatively small atomic groups within a protein; however, the movability of much larger molecular fragments, conditioned by the displacement of neighboring molecular groups, also affects the protein dynamics. In proteins, such non-local movements can take place on a very wide range of time scales that range from 10^{-9} seconds up to 10^{0} sec and even more. "Protein conformational dynamics" means precisely the non-local multiscale motions extended over many orders of time. These motions are of primary interest when studying the relationships between protein functioning and protein dynamics.

In various proteins, the interconnections between protein dynamics and protein functioning are different (see, for example, [2]). Nevertheless, there is a basic theoretical problem in this area related to the description of conformational dynamics on a range of time scales relevant to protein functioning; the

non-triviality of the problem is evident already from its formal statement. Indeed, let us view on a molecular structure as a classical system of N-particles, each with m degrees of freedom. The system states can be thought as an Euclidean space S of dimension $d = Nm$, in which a position of an imaginative point described by d-dimensional vector $\overrightarrow{R}$ is associated with a particular state of the system. Then, by introducing an energy landscape, $\Phi(\overrightarrow{R})$, over the states space S and defining the dynamic equations on the landscape $\Phi(\overrightarrow{R})$, we can study the system dynamics.

In the case of low-dimensional energy landscape $\Phi(\overrightarrow{R})$ with a few extrema, one can realize the study analytically or numerically; however, proteins are not such a case, because protein energy landscapes are extremely complex. Numerous restrictions on inter-structural movements caused by the chemical continuity of a protein macromolecule and its dense folding lead to an astronomically large number of local energy minima, "valleys" and "ridges". The protein energy landscape turns out to be so rugged that its exhaustive presentation seems impossible, even with extreme computing resources. Computer reconstructions of protein energy landscapes give either a detailed picture of protein behavior in a relatively small area of the protein conformational space, or a sketch of coarse-grained protein dynamics on the whole space of the states (see, for example, [3] and references therein). Therefore, analytical modeling of protein dynamics, which would be relevant to protein functioning, remains a challenging theoretical problem.

In this paper, we present the studies of a model of protein functioning that retains the multiscale description due to the model's design. The model is constructed within the framework of a representation of a multidimensional, rugged energy landscape by hierarchically nested basins of local minima, with an approximation of the dynamics on the basins via an ultrametric random process. This approach was developed in [4–9]. An ultrametric description of the protein operation cycle was announced earlier in [10]. We remain in a line with these ideas, yet present here a more detailed consideration of those regimes in which a protein can work. The presented results, we believe, significantly expand ideas on the regulation of enzymatic activity.

Our approach is partially contrasted to the pioneer model [11] that explicitly linked protein functioning with the protein's conformational dynamics. A set of models similar to [11] are available in later publications (see, for example, [12, 13]). We would like to emphasize that the basic idea of [11], to perceive the protein conformational dynamics as a random process propagating in a space of conformational states, seems physically reasonable; we are based on the same view. However, the approximation of stochastic protein dynamics by familiar diffusion in a well potential seems overly rough. Such an approximation deprives the protein dynamics of their multiscale character. This is crucial, because the combination of high dimensionality of the system states with multiscale stochastic dynamics can lead to the critical behavior of the system. Such expectations for proteins presume a rich diversity of dynamic modes in which the proteins can work.

The paper is organized as follows. We start from a scheme of the cyclic working of a protein, common for us and [11], and give short commentary on the

main propositions of the model [11]. Then, we set forth the ideas of the multiscale description of protein dynamics and introduce our model of the operation cycle, focusing on its particular qualities. Finally, we describe the dynamical regimes of the working cycle and give them physical interpretations.

2. Model architecture

Similar to [11], we consider a model of the cyclic binding and unbinding of a small ligand to a protein. The simplest scheme suggests two specific conformations of the protein: (E_1) refers to the equilibrated state of the bound ligand-protein system, and (E_2) refers to an unbound equilibrated state of the ligand-free protein. It is assumed that the states E_1 and E_2 are sufficiently remote from each other in their conformations. Then, the protein operation cycle proceeds as follows. Let the ligand-free protein be in the equilibrium state, E_2. The binding of a ligand takes the protein out of the equilibrium and the conformational rearrangements toward the bounded equilibrium state, E_1, start. When the ligand–protein system reaches E_1, the ligand separates from the protein and conformational rearrangements start in the opposite direction to the unbound state, E_2. Thus, the ligand binding and unbinding are driven by cyclic conformational rearrangements between two specific conformational states, E_1 and E_2.

Formally, the two-state operation cycle can be described as follows. Let B be a space of protein conformational states, $x \in B$ is a conformational state, and the bound and unbound specific states ($E_1 = x_1$ and $E_2 = x_2$) lie respectively in the domains $O_1 \subset B$ and $O_2 \subset B$, which are quite remote from each other. Let $P_1(x,t)$ be the distribution of bound proteins over the space B at time t and $P_2(x,t)$ be the distribution of unbound proteins over the same space. The total concentration of the bound and unbound proteins remains constant:

$$\int_B (P_1(x,t) + P_2(x,t))dx = 1. \tag{1}$$

The distribution $P_1(x,t)$ (and, respectively, $P_2(x,t)$) can be understood as the transition function of a random process $x(t)$ that represents the conformational dynamics of a bound (respectively, unbound) protein. That is, $P_1(x,t)$ is the conditional probability density to find a bound (respectively, unbound) protein in a state x at time t provided that at the initial time the protein was in a given state x_0. Then, the operation cycle is described by two kinetic equations of the form:

$$\begin{aligned} \frac{\partial P_1(x,t)}{\partial t} &= [\mathbf{D}_x P_1](x,t) + \lambda_1(x)P_1(x,t) - \lambda_2(x)P_2(x,t) \\ \frac{\partial P_2(x,t)}{\partial t} &= [\mathbf{D}_x P_2](x,t) - \lambda_1(x)P_1(x,t) + \lambda_2(x)P_2(x,t), \end{aligned} \tag{2}$$

where an operator $\mathbf{D}_x$ (as yet introduced formally) defines the protein conformational dynamics, and $\lambda_1(x)$ and $\lambda_2(x)$ are the rate constants of the formation and breaking of chemical bonds, respectively, between ligands and proteins. Note, that $\lambda_1(x)$ and $\lambda_2(x)$ take nonzero values only in the domains O_1 and O_2, respectively. We imply also that the ligand concentration is constant, i.e. the binding and unbinding reactions obey the monomolecular kinetics.

The operator $\mathbf{D}_x$ is a key ingredient of the model, yet its explicit definition needs a simplified representation of the protein energy landscape. In this respect, the authors of [11] made two assumptions. First, they suggested that, in view of the "reaction coordinate", the protein energy landscape can be considered as a well potential with many local minima and barriers on its walls. Then, to describe such a landscape, they introduced two characteristic scales, the smaller related to the barriers separating the local minima, and the larger refering to the well itself. These simplifications allowed them to replace a non-trivial problem of the description of stochastic dynamics on a multidimensional, rugged energy landscape with a familiar model of one-dimensional diffusion in a well potential. By these reasons, the protein conformational dynamics was described in [11] by the Fokker-Planck equation,

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2} + \frac{\partial}{\partial x}\left[\frac{1}{kT}P(x,t)\frac{\partial U(x)}{\partial x}\right], \tag{3}$$

where $P(x,t)$ is the distribution of probability density along a "conformational straight line" x, $U(x)$ is a well potential with global minima at a point x_0, D is a coefficient of conformational diffusion that depends on the "viscosity" of the conformational space, T is the temperature, and k is the Boltzmann constant. Note that the viscosity of the conformational space is caused by transitions over local barriers, and therefore the diffusion coefficient D exponentially depends on the temperature.

According to the equation (3), the protein operation cycle was described by two kinetic equations of the reaction-diffusion type:

$$\begin{aligned}\frac{\partial P_1(x,t)}{\partial t} &= D\frac{\partial^2 P_1(x,t)}{\partial x^2} + \frac{\partial}{\partial x}\left[\frac{1}{kT}P_1(x,t)\frac{\partial U_1(x)}{\partial x}\right] + \\ &+\lambda_1(x)P_1(x,t) - \lambda_2(x)P_2(x,t) \\ \frac{\partial P_2(x,t)}{\partial t} &= D\frac{\partial^2 P_2(x,t)}{\partial x^2} + \frac{\partial}{\partial x}\left[\frac{1}{kT}P_2(x,t)\frac{\partial U_2(x)}{\partial x}\right] - \\ &-\lambda_1(x)P_1(x,t) + \lambda_2(x)P_2(x,t),\end{aligned} \tag{4}$$

where $U_1(x)$ and $U_2(x)$ are the well potentials with global minima at points x_1 and x_2, positioned on the conformational straight line on some distance L from each other. In the model (4), relaxation to the bound state x_1 is governed by the potential $U_1(x)$, while the reverse relaxation to the unbound state x_2 is governed by the potential $U_2(x)$. The ligand binding and unbinding excite the protein and switch the corresponding potential.

Based on the stationary solution of the equations (4), the authors of [11] concluded that the cycle time τ is determined mainly by the diffusive moving between specific conformational states x_1 and x_2, i.e. by the ratio $L/2D$. Since the conformational diffusion coefficient D exponentially depends on T, the cycle time has the same temperature dependence and this is the only specificity in the regulation of enzymatic activity.

However, it is well known that the rates of many enzymatic reactions depend on temperature, though not exponentially or even monotonically, and often have

maxima at some temperatures; in this respect, the CO rebinding to myoglobin established in details in [14, 15] is highly indicative. It is noteworthy that the authors of [11] also mentioned these experiments and stated that their model (4) is in good agreement with the CO rebinding kinetics. We believe the opposite – the model (4) is in evident contradictions with the observations [14, 15].

In the basic experiments [14, 15], myoglobin molecules preliminarily bounded with CO were irradiated by a laser pulse, which broke the chemical bounds between the ligand and proteins. Immediately after, the kinetics of CO rebinding to myoglobin were monitored on many time scales ($10^{-7}\nabla \cdot 10^{2}$ sec) and in a wide temperature range ($300\nabla \cdot 60$K and below).

It is obvious that the CO rebinding process directly corresponds to a half of the cycle (4), so these equations should be consistent with the CO rebinding kinetics, at least quantitatively. Two outstanding features of the CO-rebinding kinetics were identified in [14, 15]. Firstly, the CO-rebinding , being limited by the conformational rearrangements of the protein, had the power-law kinetics over the wide time-window of the observations. This fact clearly indicates that the conformational dynamics have multiple scales, which unambiguously contradicts to the assumption that the barriers on the landscape can be averaged and taken into account as the effective viscosity of the conformational space. Secondly, the experiments showed that the rebinding rate grows with temperature decreasing from 300K to 200K; however, it also decreases with subsequent lowering of the temperature. Non unidirectional responses due to temperature changes are also not consistent with the model (4). Indeed, a general solution of the equation

$$\frac{\partial P(x,t)}{\partial t} = D\frac{\partial^2 P(x,t)}{\partial x^2} + \frac{\partial}{\partial x}\left[\frac{1}{kT}P(x,t)\frac{\partial U(x)}{\partial x}\right] - \lambda(x)P(x,t) \tag{5}$$

has the well known form

$$P(x,t) = \sum_{n=0}^{\infty} L(x,A_n,\gamma_n)e^{-\gamma_n Dt}, \tag{6}$$

where $L(x,...)$ is specified by functions $\lambda(x)$ and $U(x)$, while parameters A_n and γ_n are defined from the boundary and initial conditions. It is easy to see that the temperature behavior of solution (6) is determined exclusively by the temperature dependence of the conformational diffusion coefficient D, yet it is exponential in model (4), as mentioned above. Hence, there is no temperature agreement to the CO-rebinding is in the model (4).

The model suggested in [11] is not consistent with the actual features of the CO-rebinding kinetics due to its basic assumptions in its coarse description of the protein energy landscape and protein dynamics. The conclusions about the regulation of enzymatic activity, being made from so crude a description, seems to be incomplete and partly incorrect. One needs a more subtle description of protein dynamics that, above all, could hold protein dynamic's multiscale complexity.

3. Multiscale description of protein conformational dynamics

The multiscale description of protein conformational dynamics and its verifications were developed in [5, 7, 9]. Here we present the approach itself and explain the analytical notations used.

The approach is based on a hierarchical scaling of complex landscapes, just as is done in topography by means of a hierarchy of cross-sections. Applying this trick against a multidimensional, rugged landscape, we can construct a hierarchy of nested "basins" of local minima where each larger basin consists of smaller ones, each of them consists of even smaller basins, and so on (see, for example, [3]). Larger basins are separated with higher barriers, while smaller basins within larger ones are separated with lower barriers. Note that such mapping is multiscale by construction, yet it does not preserve the Euclidean metric of an original landscape.

It is natural to represent the hierarchy of nested basins by a branching tree. In such a representation, the terminal nodes $i = 1, 2, \dots$ of the tree are the local energy minima associated with the system states, and the system dynamics are interpreted in terms of random transitions between the states. Since any leaf i on the tree boundary can be parameterized by a unique branch leading from the root to the leaf, any two states differ in the divergent part of the two corresponding branches. As a consequence, any two states differ by the scale of a minimal basin that contains two given states and the metrics on the space of states obeys the strong triangle inequality $d(i,k) \leq \max(d(i,j), d(j,k))$. Thus, the mapping of a multidimensional, rugged energy landscape into a hierarchy of nested basins of local energy minima substitutes the multidimensional space of states with Euclidian metric for an ultrametric space of states. Note that the two states, being close in Euclidean metric, can turn out to be far from each other in ultrametrics, and vice versa.

The tree-like mapping looks simple enough for a regular tree of basins with fixed branching index p (see Figure 1). For any two states (leaves), i and j, there is one, and only one, minimal basin (subtree) that contains these two states. Hierarchy level $\gamma(i,j)$, $\gamma = 1, 2, \dots, \gamma_{\max}$, on which a root-vertex of the subtree lies, completely specifies the transitions between i and j; namely, it specifies the basin of $p^{\gamma(i,j)}$ states where the transition occurs, the ultrametric distance $p^{\gamma(i,j)}$ to which the transition occurs, and the highest barrier $H_{\gamma(i,j)}$ over which the transition occurs. Accordingly, the rate $w(i|j)$ of transitions between i and j is determined by the ultrametric distance between the states. By this reasoning, the transition rates are specified only by hierarchy levels $\gamma(i,j)$ and the transition matrix $\mathbf{W}$ of an *ultrametric* random walk has the characteristic block-hierarchical structure related to the basin hierarchy.

Having the transition matrix $\mathbf{W}$, one can construct the muster equation for ultrametric random walk. Let $P(i,t|i_0,0)$ be the transition probability function of the random walks, i.e. the probability to find a system in state i at time t if it was in a particular state i_0 at the initial time, $t = 0$. The transition function is a probability measure of all paths of length t which start in i_0 and end in i. For shortness, we will denote the transition probability as $P(i,t)$.

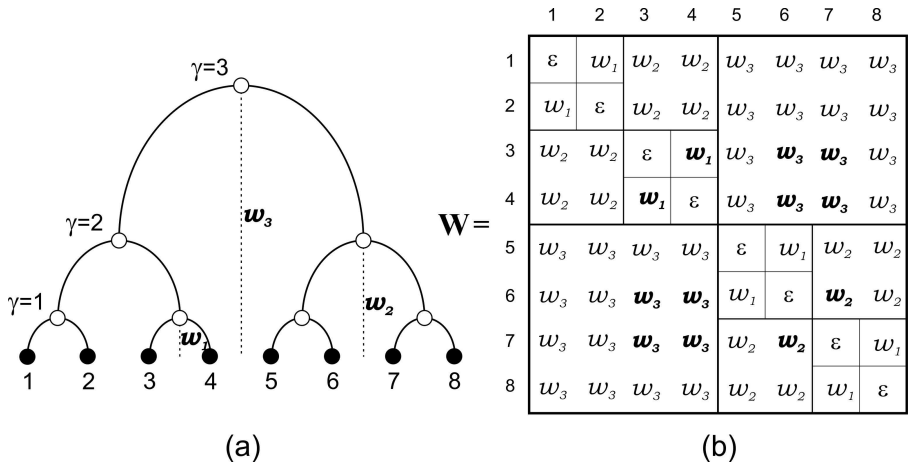


FIG. 1: (a) Regularly branching tree of basins: $\gamma = 1, 2, 3$ are the hierarchy levels; w_γ are the transition rates constants between the leaves of the tree. (b) Block-hierarchical structure of the transition matrix.

Now, let the transition probability $P(i,t)$ obeys the muster equation of the form

$$\frac{\partial P(i,t)}{\partial t} = \sum_{i \neq j} w(i|j)P(j,t) - \sum_{i \neq j} w(j|i)P(i,t) \tag{7}$$

with the initial condition $P(i,0) = \delta(i - i_0)$, where $w(i|j)$ are the transition rate constants. Assuming that the probability to escape from any basin decreases with an increasing of the basin scale, and, when the system transits in to a basin, it occupies any of its states with equal probability, we can rewrite the equation (7) as follows:

$$\frac{\partial P(i,t)}{\partial t} = \sum_{i \neq j} w(|i - j|_p)P(j,t) - \sum_{i \neq j} w(|i - j|_p)P(i,t), \tag{8}$$

where $|i - j|_p = p^{\gamma(i,j)}$ is an ultrametric distance between i and j.

The continuous limit of the discrete random walk (8) leads to the p-adic equation of ultrametric diffusion (for details see [4, 9]; for introduction in p-adic analysis see, for example, [16]). A sketch of the continuous limit is as follows. Firstly, the leaves $i = 1, 2, \ldots$ on the tree boundary are parameterized by a set of rational numbers using a mapping

$$i \longleftrightarrow x = p^{-r}\left(a_0 + a_1 p + \cdots + a_{m-1}p^{m-1}\right), \tag{9}$$

where $p \geq 2$ is a prime number, $0 \leq a_k \leq p - 1$, $k = 0, 1, \ldots, m - 1$ and r is an integer. On the rational numbers $\{x\}$ of the form (9), the p-adic norm $|\cdot|_p$ is defined as

$$|x|_p = |p^{-r}\left(a_0 + a_1 p + \cdots + a_s p^s \cdots + a_{m-1}p^{m-1}\right)|_p = p^r$$

if $a_0 = a_1 = \cdots = a_{s-1} = 0$ and $a_s \neq 0$. Then, in the limit $m \to \infty$, every element of the set $\{x\}$ can be represented as an infinite series

$$x = p^{-r} \sum_{i=0}^{\infty} a_i p^i, \tag{10}$$

that converges in the p-adic norm $|\cdot|_p$. Therefore, in the limit $m \to \infty$, the set of rational numbers $\{x\}$ becomes a compact subset B_r of the field Q_p of p-adic numbers. The limit $r \to \infty$ extends the subset B_r to the field Q_p. On the field Q_p, one can enter the ultrametric $d(x,y) = |x-y|_p$ that makes the field Q_p an ultrametric space.

Thereby, the system states x become the elements of Q_p, and the muster equation (8) transforms into a p-adic Kolmogorov-Feller equation for a homogeneous stationary Markov process on Q_p:

$$\frac{\partial}{\partial t} P(x,t) = \int_{Q_p} w\left(|x-y|_p\right) \left(P(y,t) - P(x,t)\right) d_p y, \tag{11}$$

where the rate constants of transitions are defined as

$$w\left(|x-y|_p\right) = \lim_{t' \to t} \frac{P(y,t'|x,t)}{|t'-t|}.$$

Note that the function $w(|x-y|_p)$ can have various forms. Therefore, the ultrametric diffusion equation (11) describes a family of ultrametric random processes. The particular form of $w(|x-y|_p)$ is chosen by following the reasons related to the question under the study. For protein conformational dynamics, $w(|x-y|_p)$ should conform to the CO rebinding kinetics, and such a function has the form $w(|x-y|_p) = |x-y|_p^{-(\alpha+1)}$, where $\alpha \sim T^{-1}$ (see [6]).

Thus, in our model of the protein operation cycle, the protein conformational dynamics are not described by the Fokker-Planck equation, as was done in [11]; they are described by the ultrametric diffusion equation

$$\frac{\partial P(x,t)}{\partial t} = \int_{B_r} \frac{P(y,t) - P(x,t)}{|y-x|_p^{\alpha+1}} d_p y, \tag{12}$$

where $B_r \subset Q_p$ is an ultrametric space of the protein conformational states, and $d_p y$ is the integration measure on Q_p.

Notably, the transition rates function $w(|x-y|_p) = |x-y|_p^{-(\alpha+1)}$ corresponds to *self-similar* energy landscapes, which may have a general attitude toward functional macromolecular structures like molecular machines [17, 18].

4. Basic operation modes

Let us now redefine the model (4) of the operation cycle using the equation (12) of protein conformational dynamics. Let the ultrametric ball $B_r = \left\{ x \in Q_p : \ |x|_p \leq p^r \right\}$

of radius p^r, $r \gg 1$, be the protein conformational space. Then, the operator $\mathbf{D}_x$ in the equation (2) is defined as

$$[\mathbf{D}_x P](x,t) = \int_{B_r} \frac{P(y,t) - P(x,t)}{|y-x|_p^{\alpha+1}} d_p y$$

and the operation cycle is described by the kinetic equations

$$\begin{aligned} \frac{\partial P_1(x,t)}{\partial t} &= \int_{B_r} \frac{P(y,t) - P(x,t)}{|y-x|_p^{\alpha+1}} d_p y + \lambda_1 \Omega(|x|_p) P_2(x,t) - \\ &\quad - \lambda_2 \Omega(|x-a|_p) P_1(x,t) \\ \frac{\partial P_2(x,t)}{\partial t} &= \int_{B_r} \frac{P(y,t) - P(x,t)}{|y-x|_p^{\alpha+1}} d_p y - \lambda_1 \Omega(|x|_p) P_2(x,t) + \\ &\quad + \lambda_2 \Omega(|x-a|_p) P_1(x,t). \end{aligned} \tag{13}$$

In the equations above, the distributions $P_1(x,t)$ and $P_2(x,t)$ over the conformational space B_r relate to bounded and unbounded proteins, respectively; the rate constants for the reactions of formation and breaking of chemical bonds between ligands and proteins are given by λ_1 and λ_2; and the indicator $\Omega(|z|_p)$ of an ultrametric ball of unit radius with the center $z = 0$,

$$\Omega(|z|_p) = \begin{cases} 1, & |z|_p \le 1 \\ 0, & |z|_p > 1, \end{cases}$$

specifies the binding and unbinding areas. Namely, the ultrametric ball $\Omega(|x|_p)$ with the center $x = 0$ specifies the domain O_1 where the ligands bind to proteins, while the ball $\Omega(|x-a|_p)$ with center $x = a$ specifies the domain O_2 where the ligands unbind the proteins.

The method used to solve the Cauchy problems for the equations (13) was discussed in [10], yet it is difficult to write the solution explicitly. For our purposes, however, it is sufficient to know the stationary solution $P_{1st}(x)$, as well as an estimation of the cycle time τ. The stationary solution of equations (13) is written out explicitly:

$$\begin{aligned} P_{1st}(x) &= p^{-r} \cdot \frac{\lambda_1 + \lambda_1\lambda_2 \left(I(0) - I(|a|_p) + I(|x|_p) - I(|x-a|_p) \right)}{\lambda_1 + \lambda_2 + 2\lambda_1\lambda_2 \left(I(0) - I(|a|_p) \right)} \\ P_{2st}(x) &= p^{-r} \cdot \frac{\lambda_2 + \lambda_1\lambda_2 \left(I(0) - I(|a|_p) - I(|x|_p) + I(|x-a|_p) \right)}{\lambda_1 + \lambda_2 + 2\lambda_1\lambda_2 \left(I(0) - I(|a|_p) \right)}, \end{aligned} \tag{14}$$

where

$$\begin{aligned} I(|x|_p) = \sum_{i=n}^{r-1} p^{-i} \left(1 - p^{-1}\right) \left(p^{-\alpha i} - \left(1 - p^{-1}\right) \frac{p^{-\alpha r}}{1 - p^{-\alpha-1}} \right)^{-1} - \\ - \left(p^{\alpha(1-n)} - \left(1 - p^{-1}\right) \frac{p^{-\alpha r}}{1 - p^{-\alpha-1}} \right)^{-1} p^{-n} (1 - \Omega(|x|_p)) \end{aligned} \tag{15}$$

and n is determined by the relation $|x|_p = p^n$. The stationary concentration of bounded proteins $S(\alpha, \lambda_1, \lambda_2, m)$ is directly calculated from (14):

$$S(\alpha, \lambda_1, \lambda_2, m) = \int_{B_r} P_{1st}(x) d_p x = \frac{\lambda_1 + \lambda_1 \lambda_2 \cdot [I(0) - I(p^m)]}{\lambda_1 + \lambda_2 + 2\lambda_1 \lambda_2 \cdot [I(0) - I(p^m)]}. \qquad (16)$$

The sensitivity of $S(\alpha, \lambda_1, \lambda_2, m)$ to variations of the parameters is our main interest.

The cycle time, τ, is determined by the slowest stages. In some cases, they may be the stages of formation or breaking of the chemical bonds between ligands and proteins; such a mode is associated with the kinetic control. In other cases, the protein conformational rearrangements may be the slowest stage. Since the protein dynamics are interpreted as a stochastic process propagating in the protein conformational space, this mode can be associated with the diffusion control. In this respect, the features of ultrametric diffusion become important for specifying the operation modes.

To clarify how ultrametric diffusion propagates, let us consider the solution $P(x,t)$ of the Cauchy problem for equation (12) with an initial distribution $P(x,0) = \Omega(|x|_p)$. The solution has the form (see [4–6]):

$$P(x,t) = (1-p^{-1})|x|_p^{-1} \sum_{\gamma=0}^{+\infty} p^{-\gamma} \Omega\left(\frac{p^{-\gamma}}{|x|_p}\right) \exp\left\{\frac{p^{-\alpha\gamma}}{|x|_p^{\alpha}} t\right\} - \\ - |x|_p^{-1} \Omega\left(\frac{p}{|x|_p}\right) \exp\left\{\frac{p^{\alpha}}{|x|_p^{\alpha}} t\right\}. \qquad (17)$$

Knowing the distribution $P(x,t)$, we can calculate an average ultrametric distance $\delta(t)$,

$$\delta(t) = \int_{Q_p} |x|_p P(x,t) d_p x, \qquad (18)$$

for which a diffusion front moves by the time t. Using the solution (17), it is not difficult to see that the average distance $\delta(t)$ is finite only if $\alpha > 1$. For $\alpha \leq 1$, the integral (18) diverges. Therefore, unlike the familiar diffusion, we can speak about the front of ultrametric diffusion only if $\alpha > 1$, i.e., when the diffusion propagates rather "slowly". The diffusion mode corresponds exactly to such conditions. In such regimes, an estimation of the cycle time τ can be made from the simple expression $\delta(\tau) = p^m$.

In contrast to this case, if $1 \geq \alpha > 0$, ultrametric diffusion is so fast that the distribution $P(x,t)$ almost immediately becomes not small at all points of the conformational space [8]. In other words, the diffusion front is delocalized. Note that the ultrametric diffusion delocalizes sharply, exactly at the critical point $\alpha = 1$. If the space is finite, as in our model (13), the typical trajectories first get far from an initial state and only then do they come back. For reaction-diffusion models, this condition physically responds to the rapid mixing of the diffusion volume. Note that B_r is much larger than the binding-unbinding domains O_1 and O_2. Therefore, the rapid mixing of a large conformational volume makes

the distributions $P_{1st}(x)$ and $P_{2st}(x)$ almost homogeneous on B_r. Hence, the concentrations of bounded and unbounded proteins become practically equal

$$\int_{B_r} P_{1st}(x)d_px \approx \int_{B_r} P_{2st}(x)d_px$$

even if $\lambda_1 \neq \lambda_2$. Some interesting details of this regime are discussed in the nest section.

Therefore, there is a critical value of α, at which the ultrametric diffusion delocalizes. This fact suggests that the transition from the diffusion mode to the kinetic mode may occur in a rather narrow area of the parameter values. Compared with the model (4), this is one more specificity of the model (13).

5. Cycle control

The response of a stationary concentration of bounded protein to changes of parameters α, λ_1, λ_2, and m allow us to get an idea of the regulations of enzymatic activity through the modifications of proteins or effects on the environment. Indeed, parameter α specifies the transition rates of ultrametric diffusion, $w(x|y) = |x-y|_p^{\alpha+1}$ (see equation (12)). In this sense, α reflects the protein conformation mobility: larger values of α result in lower conformation mobility. Changes in conformation mobility can be achieved by variations of the temperature, viscosity, or ionic composition of the solution, incorporation of a protein into a cell membrane, or the loading of a protein by molecular compounds.

Changing the rate constants of the formation and breaking of chemical bonds between proteins and ligands, λ_1 and λ_2, can occur due to familiar temperature dependence or, for instance, chemical modifications of ligands or protein binding sites. Finally, the parameter m, which sets an ultrametric distance between the domains O_1 and O_2, regulates the minimal depth of the conformational rearrangements needed for working.

Thus, all model parameters have transparent physical interpretations.

Now, let us consider the sensitivity of stationary states $S(\alpha, \lambda_1, \lambda_2)$ to variations of the model parameters. First, we discuss the sensitivity to protein conformational mobility. As noted above, there are two regimes separated by a critical value of α, at which the front of ultrametric diffusion is delocalized. These two regimes can be seen from behavior of the term $\Delta I = I(0) - I(p^m)$ in the expression (16) (see Figure 2).

When the diffusion front is localized ($\alpha > 1$), the term $\Delta I(m,\alpha)$ increases with the growth of α. For sufficiently large α, $\Delta I(m,\alpha) \sim e^{k(m)\alpha}$, where $k(m)$ depends on m linearly. Indeed, using the expression (15) and taking into account that $(1-\Omega(p^m)) = 1$ for $m \geq 1$, one can write

$$\Delta I(m,\alpha) = \sum_{i=0}^{m-1} \frac{p^{-i} - p^{-i-1} - p^{-\alpha-i-1} + p^{-\alpha-i-2}}{p^{-\alpha i} - p^{-\alpha i-\alpha-1} - p^{-\alpha r} + p^{-1-\alpha r}} -$$
$$- \frac{1-p^{-\alpha-1}}{p^{\alpha} - p^{-1} - p^{-\alpha r} + p^{-1-\alpha r}} + \frac{p^{-m} - p^{-\alpha-m-1}}{p^{\alpha-\alpha m} - p^{-1-\alpha m} - p^{-\alpha r} + p^{-1-\alpha r}}. \qquad (19)$$

For large α we have

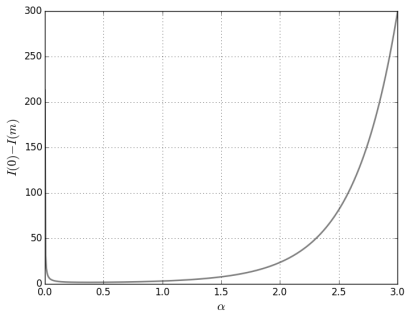


FIG. 2: Dependence of the term $\Delta I = I(0) - I(p^m)$ in the expression (16) from α.

$$\Delta I(m,\alpha) \overset{\alpha\gg 1}{\longrightarrow} p^{(m-1)(\alpha-1)}.$$

For $\alpha \to 0$, $\Delta I(m,\alpha)$ also diverges:

$$\Delta I(m,\alpha) \overset{\alpha\to 0}{\longrightarrow} +\infty.$$

Between these extremal values of α, the term $\Delta I(m,\alpha)$ slowly increases with growth of α. When $\alpha < 1$, i.e. the diffusion front is delocalized, the term $\Delta I(m,\alpha)$ is small, and we obtain from the expression (16)

$$S(\alpha,\lambda_1,\lambda_2) \approx \frac{\lambda_1}{\lambda_1+\lambda_2}.$$

This is the kinetic mode: the stationary state S is not sensitive to α, yet it is sensitive to λ_1 and λ_2. In particular, if $\lambda_1 = \lambda_2$, the stationary state is symmetric, $S(\alpha,\lambda_1,\lambda_2,m) = 0,5$.

The term $\Delta I(m,\alpha)$ remains small until $\alpha \approx 1$. Only when $\alpha > 1$, i.e. the diffusion front localizes, does the contribution of $\Delta I(m\alpha)$ in to $S(\alpha,\lambda_1,\lambda_2,m)$ become significant. For rather large α, the sensitivity to λ_1 and λ_2 is lost, the operation cycle proceeds in the diffusion mode, and the stationary state S restores the symmetry even if $\lambda_1 \neq \lambda_2$. The symmetry takes place because in this regime, the cycle operation is limited by ultrametric diffusion, yet the times of forward and reverse diffusion between specific states O_1 and O_2 are the same. Note that in the diffusion mode, the scale of the conformational basin where the cycle operates is determined by the parameter m. The largest part of protein conformational space beyond this basin remains "dark".

The fact that $\Delta I(m,\alpha)$ diverges in the "high temperature limit" ($\alpha \to 0$) is also important for the cycle operation. When the protein conformation mobility is high, the contribution of $\Delta I(m,\alpha)$ into expression (16) turns out to be dominant and the stationary state S becomes nearly symmetric for any reasonable λ_1 and λ_2.

Thus, the cycle operation loses sensitivity to λ_1 and λ_2 at low and high conformation mobility (see Figure 3); however, the difference between these two

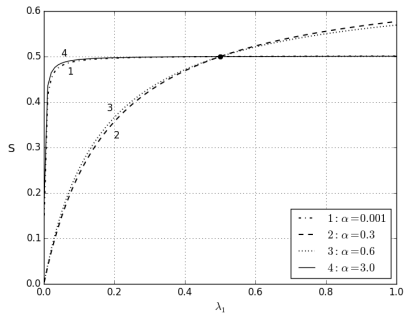


FIG. 3: Sensitivity of stationary concentration S to changes in λ_1 ($\lambda_2 = 0,5$): 1 – the conformational mixing mode; 2 and 3 – the kinetic mode; 4 – the diffusion mode

.

regimes needs to be understood. At low conformational mobility, the ultrametric diffusion propagates slowly and limits the cycle operation. In contrast, at high conformational mobility the ultrametric diffusion quickly spreads over the conformational space and the protein conformational states are rapidly mixed. For sufficiently small α, the mixing can cover large parts of the conformational space. In other words, a very large part of the conformational space that has been still remained "dark", starts to play an important role. In the diffusion delocalized regime, when the conformational mobility increases, initially, the majority of the large conformational space B_r remains "dark" and the cycle operates in the kinetic mode: the cycle is sensitive to variations in the binding and unbinding constants. However, as far as the conformational mixing expands on the majority of the conformational space, the cycle loses its sensitivity to these parameters and restores the symmetric state.

Let us now outline how the cycle operation depends on λ_1 and λ_2 in basic modes. Figures 4 and 5 represent the stationary landscapes $S(\lambda_1, \lambda_2)$ for the diffusion mode ($\alpha = 3$) and for the kinetic mode ($\alpha = 0.75$), correspondingly. As noted above, if $\lambda_1 = \lambda_2$, the stationary state is always symmetric; it is a kind of a "dead center" of the cycle.

In the diffusion mode ($\alpha = 3$, Figure 4), the ratio of λ_1 to λ_2 has no significant impact on stationary concentration S, except in the cases when these parameters are so small that they lock the cycle. Everywhere beyond the locking, the cycle operates in the diffusion mode and is close to the symmetric state.

When the diffusion front is delocalized ($\alpha = 0.75$, Figure 5), the cycle operates in kinetic mode and it is sensitive to the λ_1 and λ_2. If the conformational diffusion is not overly fast (α is not too small), sensitivity to α is also preserved. However, if the diffusion is so fast that the conformational mixing covers a large part of the conformational space, the cycle loses sensitivity to λ_1 and λ_2 and operates in the symmetric regime.

The dependence $S(\alpha)$ at a fixed ratio of λ_1 and λ_2 shown on Figure 6 clearly illustrates such behaviour. This figure also demonstrates that the regulation of

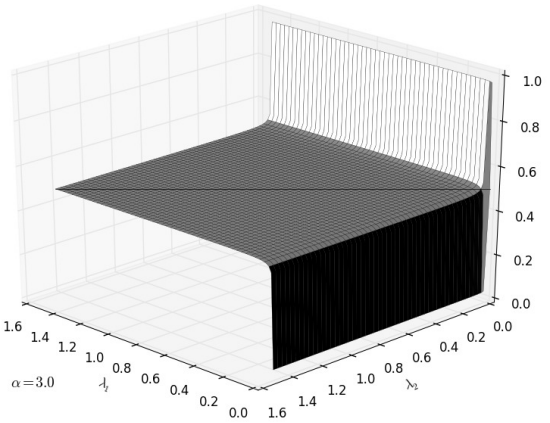


FIG. 4: The landscape of enzymatic activity at low conformational mobility: $\alpha = 3$. Here $p = 2,\ r = 7,\ m = 5$

the cycle operation is suitable at a narrow interval slightly above the diffusion delocalization point where the cycle operation is the most sensitive to parametric control.

6. Conclusion

Enzymatic activity indeed depends on external conditions in a non-trivial way. In some cases, enzymatic activity changes considerably even at small variations of temperature, solution viscosity, or the solutions's ionic composition. In other cases, enzymatic activity is not sensitive to such changes at all. For some proteins, the enzymatic activity increases monotonically with increasing temperature; for others, it decreases monotonically, and for the third type it has a maximum in a rather narrow region of the parameter values. Therefore, changing one parameter, e.g., lowering the temperature, can effect enzymatic activity in opposite ways for different proteins: in one case, enzymatic activity rises, while in others it falls.

Such different reactions to qualitatively identical impacts formed a widespread opinion that the regulation is protein–specific and could vary in opposite directions. Relatively simple low-dimensional models of the operation cycle, such as the one proposed in [11], reflect this point of view. Oppositely–directed responses simulated by unidirectional effects do not support by such models: such phenomena look "mysterious" and may be excused only by protein individualities.

In this paper, we show that the contradictory behavior of enzymatic activity, in fact, is a manifestation of the complexity of the protein energy landscape, an ultrametric models just reflect the complexity of protein dynamics. They hold both the multi-scale dynamics and an exponentially large bulk of the space of protein conformational states. It is important to emphasize that the ultrametric

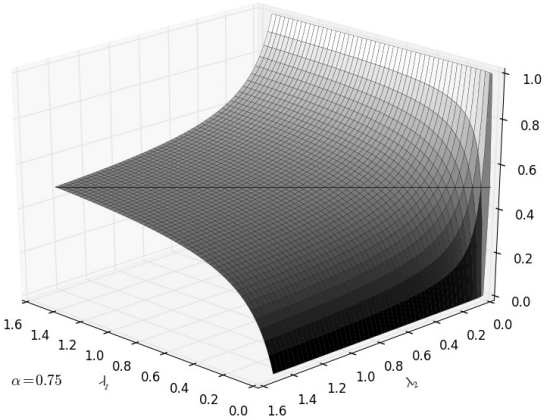


FIG. 5: The landscape of enzymatic activity at relatively high conformational mobility: $\alpha = 0.75$;other parameters are the same as those for Figure 4

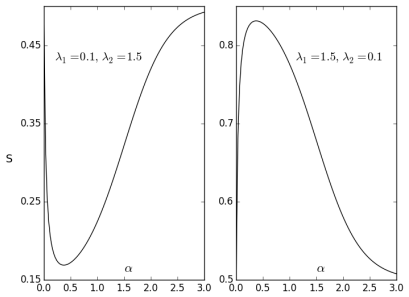


FIG. 6: Stationary concentration S dependence from α near the critical point of diffusion front's delocalization $\alpha = 1$. Values of $\lambda_1 \neq \lambda_2$ are fixed.

description of protein conformational dynamics developed in [4–10] completely agrees with the actually–observed kinetics of enzymatic reactions.

The ultrametric model of the protein operation cycle allows us to see many possibilities for the regulation of enzymatic activity that are hidden for standard low-dimensional models of the reaction-diffusion type. The most important specificity of the ultrametric models is the phenomena of conformational delocalization. The critical condition of delocalization establishes a kind of a borderline between two main modes of the cycle operation the diffusion mode, which is set up at low conformational mobility, and the kinetic mode at high conformational mobility. Because ultrametric diffusion abruptly delocalizes at the threshold mobility, the transition between the diffusion mode and the kinetic mode is carried out in a narrow area of parameters. At that area, the sensitivity of the operating cycle to variations of external conditions is high.

Another important feature of the protein operation cycle appears due to very large capacity of the protein conformational space. When the protein conformational mobility is rather low, the largest part of the protein conformational space remains ”dark“. The protein operates on a relatively small basin of conformational states and the operation cycle is sensitive to changes in external conditions. When the conformational mobility is high, the protein operates in an extremely large conformational space. At that time, the operation cycle loses sensitivity to changes in external conditions. As a result, unidirectional changes in external conditions, e.g. the rise in temperature, may affect enzymatic activity in opposite directions – all depending on the mode in which the cycle operates.

Thus, the multiscale conformational dynamics, in conjunction with extremely large bulk of the conformational space, may play a special role in regulation of enzymatic activity. The ultrametric model of the protein operation cycle reflects these important features and highlights a much richer picture, in which the different responses to similar influences is a manifestation of the complexity of the protein energy landscape and protein conformational dynamics.

References

1. Blumenfeld, L. A.: Problems of Biological Physics. Springer-Verlag, Berlin Heidelberg New York. (1981)
2. Nelson, D. L. and Cox, M. M.: Lehninger Principles of Biochemistry, 6th Edition. W.H. Freeman Publisher. (2012)
3. Wales, D.: Energy Landscapes. Applications to Clusters, Biomolecules and Glasses. Cambridge University Press, Cambridge, UK. (2004)
4. Avetisov, V. A. and Bikulov, A. Kh. and Kozyrev, S. V.: Application of p-Adic Analysis to Model of Breaking of Replica Symmetry. In Journal of Physics A: Mathematical and General, Vol. 32, 8785–8791. (1999)
5. Avetisov, V. A. and Bikulov, A. Kh. and Kozyrev, S. V. and Osipov, V. A.: p-Adic Models of Ultrametric Diffusion Constrained by Hierarchical Energy Landscapes. In Journal of Physics A: Mathematical and General, Vol. 35, 177–189. (2002)
6. Avetisov, V. A. and Bikulov, A. Kh. and Kozyrev, S. V. and Osipov V. A.: p-Adic Description of Characteristic Relaxation in Complex Systems. In Journal of Physics A: Mathematical and General, Vol. 35, 4239–4246. (2003)
7. Avetisov, V. A. and Bikulov, A. Kh.: Protein Ultrametricity and Spectral Diffusion in Deeply Frozen Proteins. In Biophysical Reviews and Letters, Vol. 3, 387–396. (2008)

8. Avetisov, V. A. and Bikulov, A. Kh. and Zubarev, A.P.: First Passage Time Distribution and the Number of Returns for Ultrametric Random Walks. In Journal of Physics A: Mathematical and Theoretical, Vol. 42, 08503–08521. (2009)
9. Avetisov, V. A. and Bikulov, A. Kh. and Zubarev, A. P.: Ultrametric Random Walk and Dynamics of Protein Molecules. In Proceedings of the Steklov Institute of Mathematics, Vol. 285, 3–25. (2014)
10. Avetisov, V. A. and Bikulov, A. Kh. and Zubarev, A. P.: On Mathematical Modeling of Molecular "Nanomachines". In Journal of Samara State Technical University, Physical and Mathematic Sciences Series, No. 1(22), 9–15. (2011)
11. Shaitan, K. V. and Rubin, A. B.: Conformational Dynamics of Proteins and Simplest Molecular "Machines". In Biophysics, Vol. 27, 386–340. (1982)
12. Chowdhury, D.: Stochastic Mechano-Chemical Kinetics of Molecular Motors: A Multidisciplinary Enterprise from a Physicists Perspective. In Physics Reports, Vol. 529, 1-197. (2013)
13. Chowdhury, D.: Modeling Stochastic Kinetics of Molecular Machines at Multiple Levels: From Molecules to Modules. In Biophysical Journal, Vol. 104, 2331-2341. (2013)
14. Ansary, A. and Berendzen, J. and Bowne, S. F. and Frauenfelder, H. and Iben, I. E. T. and Sauke, T. B. and Shyamsunder, E. and Young, R. D.: Protein States and Proteinquakes. In Proceedings of the National Academy of Sciences of the USA, Vol. 82, 5000–5003. (1985)
15. Steinbach, P. J. and Ansari, A. and Berendzen, J. and Braunstein, D. and Chu, K. and Cowen, B. and Ehrernstein, D. and Frauenfelder, G. and Johnson, J. B. and Lamb, D. C. and Luck, S. and Nienhaus, G. U. and Orinos, P. and Phillip, R. and Xie, A. and Young, R.: Ligand Binding to Heme Proteins: Connections between Dynamica and Function. In Biochemistry, Vol. 30, 3988–4001. (1991)
16. Vladimirov, V. S. and Volovich, I. V. and Zelenov, E. I.: p-Adic Analysis and Mathematical Physics. Series on Soviet and East European Mathematics, Vol. 1. World Scientific, Singapure. (1994)
17. Avetisov, V. A. and Ivanov, V. A. and Meshkov, D. A. and Nechaev, S. K.: Fractal Globules: a New Approach to Artificial Molecular Machines. In Biophysical Journal, Vol. 107, 2361–2368. (2014)
18. Avetisov, V. A and Nechaev, S. K.: Fractal Polymer Globules: A New Insight on Prebiological Evolution. In Geochemistry International, Vol. 52, No. 13, 1235–1242. (2014)

Radiation Induced Dysfunctions in the Working Memory Performance Studied by Neural Network Modeling

A.N. Bugay, G.F. Aru, E.B. Dushanov, and A.Yu. Parkhomenko

Joint Institute for Nuclear Research, Joliot-Curie 6, 141980 Dubna, Russia
bugay_aleksandr@mail.ru

Abstract. Biophysical conductance-based neural network model of working memory is proposed for study of radiation-induced impairments. The model correctly represents generation of spatially ordered structures with high cell activity emerge in the modeled brain's region. Radiation-induced changes in the synaptic receptor number and ion channel conductivities were evaluated on the basis of experimental data. In the course of calculations, an absorbed dose threshold was found, above which the stability of the time-space structures specific for the given network is lost.

Keywords: biological neural network, radiation damage

1. Introduction

The synchronization of neuronal activity within a specific network is required for cognitive performance. Normal performance of neural network may be disturbed by various external factors. Among them galactic cosmic radiation remains one of the poorly studied while providing a potential risk for central nervous system in long-term space travel [1, 2]. When evaluating the risk associated with exposure to galactic cosmic rays heavy nuclei during an interplanetary flight, the possible development of the cosmonauts central nervous system (CNS) disorders should be taken into account. NASA estimations show that beyond the Earth's magnetosphere a square centimeter is crossed by about 160 heavy charged particles with the nuclear charge $Z > 20$ during 24 hours. It has been calculated that during a three-year manned interplanetary flight, from 7 to 13 % of the CNS neurons can be exposed to high-energy iron ions, and up to 46%, to particles with $Z > 15$ [2].

In ground-based experiments, exposure to heavy ion radiation at doses matching the real fluxes of galactic iron nuclei during a Mars mission induces pronounced CNS dysfunctions [3]. Their symptoms include expressed spatial orientation disorders and suppression of cognitive functions, which is linked with damage to the synaptic transmission mechanisms [4–6], membrane ion channels [7–9].

In order to have predictive value for risks, biological pathways and their outputs need to be organized into mathematical models. Development of mathematical models for neural networks and structures seems to be an extremely important part in such research. Biological neural network simulation have been

applied recently for the quantification of related phenomena in hippocampus [10]. This model was used to compare predicted hippocampal CA1 region network firing statistics using input parameters from proton-irradiated versus control mice.

In present work we will study radiation dysfunction of neural activity in the prefrontal cortex that is responsible for short-term retention of information about the object (working memory).

2. Model Neural Network

Working memory is the ability to transiently hold and manipulate goal-related information to guide forthcoming actions. The prefrontal cortex (PFC) is the brain structure most closely linked to working memory. PFC neurons show elevated persistent activity [11] during delayed reaction tasks, when information derived from a briefly presented cue must be held in memory during a delay period to guide a forthcoming response. The activity is grouped within so called memory fields related to selected object.

Our approach is based on established models [12–14] representing common view that selective persistent neuron activity mechanism is maintained by recurrent excitation within cell assemblies. The network is composed of $N_P = 144$ pyramidal cells (excitatory population) and $N_I = 36$ interneurons (inhibitory population) according to typical architecture of a cortical module. Single compartment conductance-based neurons in the model are connected with each other by spatially structured synaptic contacts with three types of receptors. The equations for the membrane potential of neurons are as follows:

$$C\frac{dV_p}{dt} = -I_{mem,p} - I_{syn,p} - I_{noise} - I_{ext}, \tag{1}$$

$$C\frac{dV_i}{dt} = -I_{mem,i} - I_{syn,i} - I_{noise}, \tag{2}$$

where V_p and V_i are membrane potentials for pyramidal cells and interneurons, respectively, indexes $p = 1...N_P$, $i = 1...N_I$ correspond to cell number, the membrane capacitance is C = 1 nF. All voltages in the model are given in mV, and time units are given in ms.

The transmembrane ionic currents are given as follows:

$$I_{mem,p} = I_{leak} + I_{Na} + I_{Kdr} + I_{Kahp}, \tag{3}$$

$$I_{mem,i} = I_{leak} + I_{Na} + I_{Kdr}. \tag{4}$$

The leak current I_{leak} is given in common form

$$I_{leak} = g_L(V - V_L), \tag{5}$$

where maximum leakage conductance is $g_L = 0.3$nS, and corresponding rest potential is $V_L = 10.59$mV.

The expression for sodium current I_{Na} is analogous to Hodgkin-Huxley model

$$I_{Na} = g_{Na}m^3h(V - V_{Na}), \tag{6}$$

where maximum conductance is $g_{Na} = 120$nS, and reversal potential is $V_{Na} = 115$mV. The dynamics of gating variables is defined in standard form

$$\frac{dm}{dt} = a_m(V)(1-m) - b_m(V)m, \tag{7}$$

$$\frac{dh}{dt} = a_h(V)(1-h) - b_h(V)h, \tag{8}$$

where

$$a_m = \frac{0.1(25-V)}{\exp[0.1(25-V)]-1}, \qquad b_m = 4\exp[-V/18],$$

$$a_h = 0.07\exp[-V/20], \qquad b_h = \frac{1}{\exp[0.1(30-V)]+1}.$$

Potassium fast delayed rectifier current I_{Kdr} is also taken according to Hodgkin-Huxley model

$$I_{Kdr} = g_K n^4 (V - V_K). \tag{9}$$

The maximum conductance is $g_{Kdr} = 36$nS, and reversal potential is $V_{Na} = -12$mV. The dynamics of gating variable is given by

$$\frac{dn}{dt} = a_n(V)(1-n) - b_n(V)n, \tag{10}$$

where

$$a_n = \frac{0.1(25-V)}{\exp[0.1(25-V)]-1}, \qquad b_n = 4\exp[-V/18].$$

Slow Ca^{2+}-activated potassium current I_{Kahp}, which mediates a slow after-hyper-polarization (AHP) and spike frequency adaptation is given as follows:

$$I_{Kahp} = \frac{g_{ahp} q (V - V_K)}{1+q}. \tag{11}$$

Gating variable q is controlled by Ca^{2+}:

$$\frac{dq}{dt} = a_q(V)(V - V_{Ca}) - b_q(V)q, \tag{12}$$

where $g_{Kahp} = 0.01$nS, $g_{Ca} = 0.01$nS, $V_{Ca} = 185$mV, and

$$a_q = -\frac{0.002 g_{Ca}}{1+\exp[-(V-45)/4]}, \qquad b_q = 1/80.$$

Synaptic input to each cell includes three components:

$$I_{syn} = I_{NMDA} + I_{AMPA} + I_{GABA}. \tag{13}$$

First two types of synaptic connections coming from pyramidal cells are excitatory with glutamate as transmitter. Postsynaptic current formed by N-methyl-D-aspartate (NMDA) receptors has nonlinear voltage dependence, which is referred to the magnesium block

$$I_{NMDA,k} = \frac{g_{NMDA,x}(V_k - V_E)}{1 + 2.975\exp[-0.062(V_k - V_E)]} \sum_{j=1}^{N_p} W_x(j,k) S_{NMDA,j} \tag{14}$$

where

$$\frac{dS_{NMDA,j}}{dt} = K_n(V_j)(1 - S_{NMDA,j}) - S_{NMDA,j}/\tau_n, \tag{15}$$

$$K_n = \frac{0.66}{1 + \exp[-(V-30)/2]}, \tau_n = 40.$$

Another excitatory current formed by a-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA) receptors is linear with respect to postsynaptic voltage:

$$I_{AMPA,k} = g_{AMPA,x}(V_k - V_E) \sum_{j=1}^{N_p} W_x(j,k) S_{AMPA,j} \tag{16}$$

where

$$\frac{dS_{AMPA,j}}{dt} = K_a(V_j)(1 - S_{AMPA,j}) - S_{AMPA,j}/\tau_a, \tag{17}$$

$$K_a = \frac{22}{1 + \exp[-(V-30)/2]}, \tau_a = 1.5.$$

Inhibitory synaptic current formed by gamma-aminobutyric acid type-A (GABA) receptors is given by the following expression:

$$I_{GABA,k} = g_{GABA,x}(V_k - V_I) \sum_{j=1}^{N_i} W_x(j,k) S_{GABA,j} \tag{18}$$

where

$$\frac{dS_{GABA,j}}{dt} = K_g(V_j)(1 - S_{GABA,j}) - S_{GABA,j}/\tau_g, \tag{19}$$

$$K_g = \frac{20}{1 + \exp[-(V-30)/2]}, \tau_g = 5.$$

The reversal potentials are $V_E = 65$mV for excitatory synapses, and $V_I = -12$mV for inhibitory synapses, respectively.

Spatial distribution of synaptic contacts is defined by weights functions:

$$W_x(j,k) = \frac{1}{2\sigma_x} \exp(-|j-k|/\sigma_x). \tag{20}$$

Here indexes 'x' denote the type of connection. For $p \to p$ connections $g_{AMPA} = 1.82$nS, $g_{NMDA} = 5.89$nS, $\sigma = 3.2$. For $p \to i$ connections $g_{AMPA} = 1.46$nS, $g_{NMDA} = 4.64$nS, $\sigma = 1.2$. For $i \to p$ connections $g_{GABA} = 5.63$nS, $\sigma = 4.8$. For $i \to i$ connections $g_{GABA} = 4.38$nS, $\sigma = 1.2$.

Noise was implemented as single spikes arriving at average frequency on a random cell according to a uniform distribution.

Constructed model preserves basic properties common for normal working memory performance. Typical simulation result of neural network activity after presented cue is shown on Fig.1. Presented 100 ms cue contained four parts evenly divided across the whole pyramidal cell population. After end of cue four groups of stimulated pyramidal cells maintain their firing, without spreading their activity to other cells. Interneurons also respond a spike series after after external stimulation.

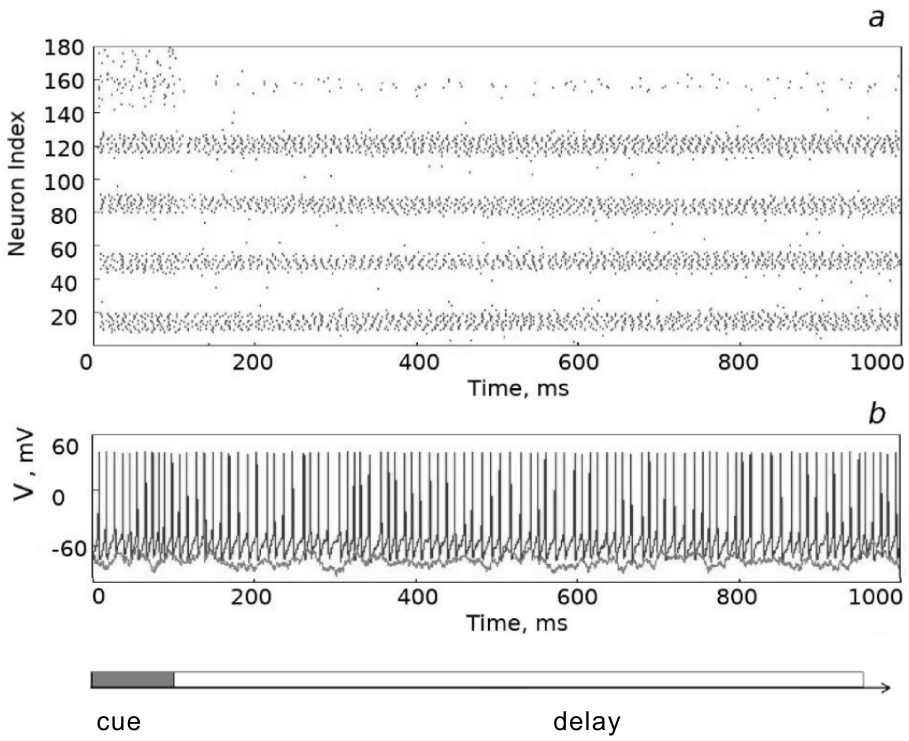


FIG. 1: The activity of whole population containing 144 pyramidal cells (from 1 to 144) and 36 interneurons (from 145 to 180). a) Four groups of stimulated pyramidal cells (black bars correspond to spikes of action potentials) do not spread their activity to other cells. b) Single neurons in the activated column maintain their firing after external stimulation ends.

3. Radiation Induced Effects and Network Performance

Early studies through the 1960s using electrons, X-rays and gamma rays showed altered electrophysiology in neurons after irradiation [15]. It was show that irradiation causes an increased permeability of the membrane (leakage) and a net loss of potassium ions [7]. Analogous results had been reported that radiation directly alters the properties of sodium ion channels on membrane [8, 9]. Recent findings suggested that intrinsic nerve properties were relatively resistant, but that synapses might be more sensitive targets. In the latter case there were observed reduced presynaptic release of glutamate [5] and GABA [6] and also decreased abundance of glutamate receptors in synaptosomes [4, 5].

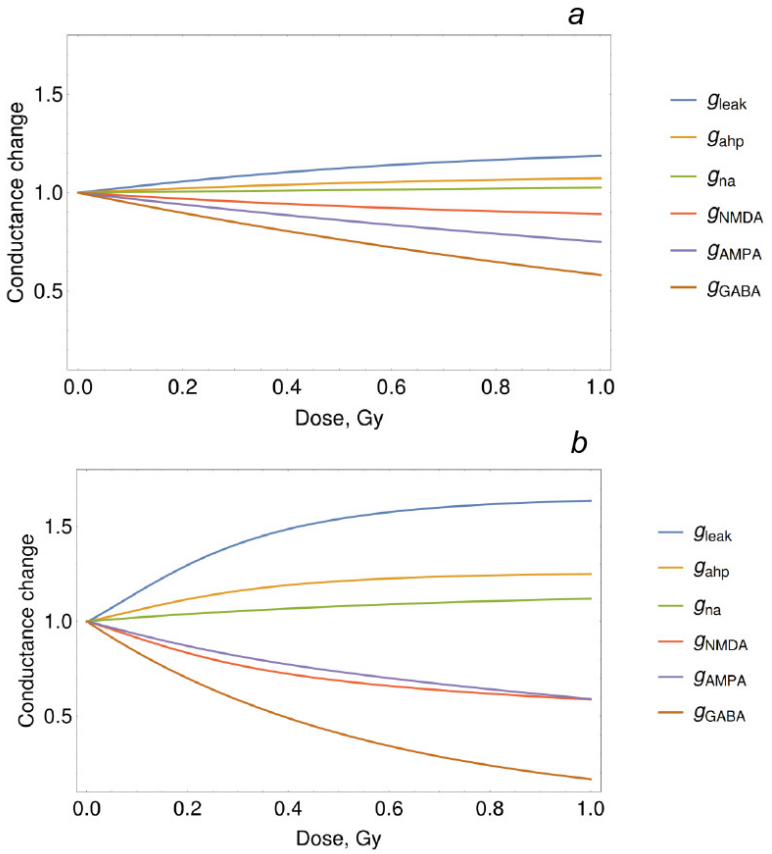


FIG. 2: Relative radiation-induced conductance change estimated for 150 MeV/n protons (a) and for 600 MeV/n Fe ions.

A complex number of radiation alteration may lead to disruption of normal neuron electrophysiology. The reason for mentioned alterations in not firmly established yet. Along with direct damage to sensitive structures from traversing

particle tracks there may be secondary effects such as oxidative stress caused by reactive oxygen species (ROS). Critical regulatory sites for neuronal activity are receptor-gated ion channels, and recent evidence suggests that multiple channels are regulated by their redox status. Both GABA [16] and glutamate [17] receptors were shown to be susceptible to oxidation, which resulted in changes in ion conductance and channel opening probability.

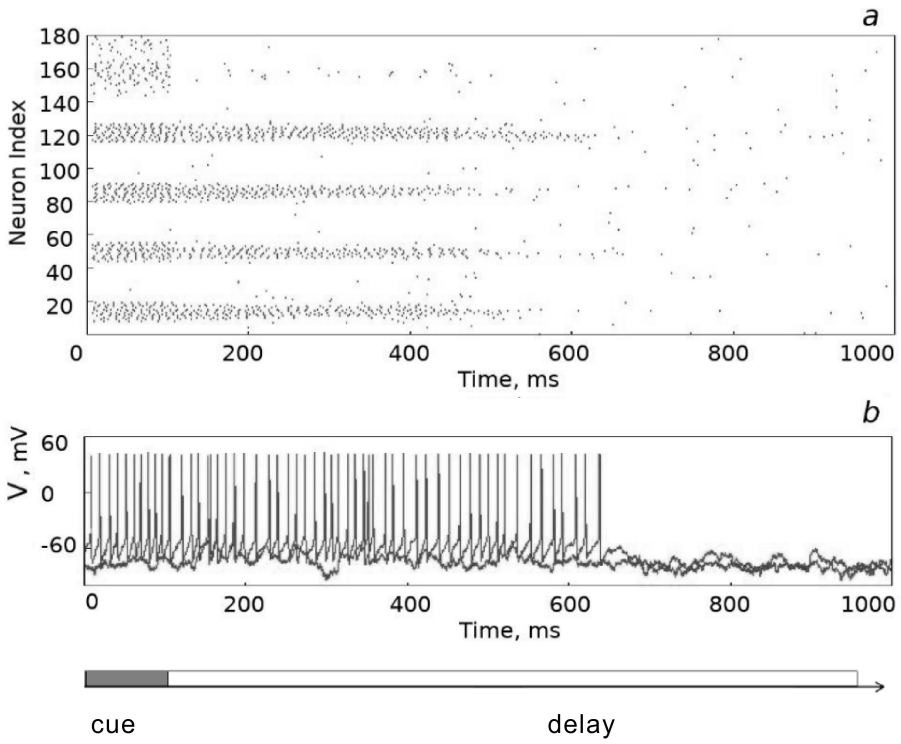


FIG. 3: Simulation of neural network activity after irradiation of 0.5 Gy 600 MeV/n iron ions. The initial conditions are the same as in control shown in Fig.1. All four groups of stimulated pyramidal cells have lost their ability to maintain firing after external stimulation.

Another secondary effect is hiden in perturbation of brain neurochemistry. The concentrations of monoamines and their metabolites in different brain areas including the prefrontal cortex were shown to be affected by ionizing radiation [18, 19]. It is suggested that the effect of dopamine on WM performance is mediated by D1 receptors. Dopaminergic modulation via the D1 receptor affects transmission through the AMPA and NMDA receptors, persistent sodium current, the Ca^{2+}-dependent potassium current I_{Kahp}, and the spontaneous activity of interneurons in the PFC [14]. It was shown that increase or decrease

of domamine level far from optimal value strongly affects the stability and the capacity of working memory [14].

At the current level of knowledge it is not possible to develop self-consistent theoretical model of radiation-induced damage to critical sites of neurons. Existing models of charged particle interaction with neurons are only available to count energy deposition events in critical sites of neural cell [20, 21]. Therefore, we will follow the approach introduced in [10], which is based on the usage of experimentally determined change in parameters of neural network. In our case required set of model parameters should be estimated from existing experimental data. Relative conductance change after the irradiation can be expressed in the following general form

$$\bar{g}_x = g_x R_x[D] M_x[Z_{da}(D)]. \tag{21}$$

Here index 'x' stands for the type of ionic or synaptic conductance, $\bar{g}_x$ is the modified conductance with respect to absorbed radiation dose D.

Function R describes direct damages to ionic channels as a result of energy deposition or oxidation. After sufficiently long time oxidation is supposed to prevail over direct damage, therefore it mostly defines dose dependence of R. An analytic expression for production of ROS with respect to dose and particle type is unknown. Therefore, we have taken interpolated data from experimental curves in [22] and adopted them to known values conductivity changes taken from [4–9]. Further we will compare effects of light particles (150 MeV/n protons) and heavy particles (600 MeV/n ^{56}Fe ions), which have completely different linear energy transfer (LET) and relative biological efficiency. Most relevant experimental works on synapses and ionic channels are related to this choice.

Function M describes selective dopaminergic modulation of ionic channels and receptors according to interpolated curve according to experimentally estimated levels of dopamine concentration Z_{da} [18, 19]. Analytic dependence of M is chosen according to [14]:

$$M_x = c_x \left(1 + \frac{d_x}{1 + \exp[(a_x - Z_{da})/b_x]} \right), \tag{22}$$

where parameters a_x, b_x, c_x, d_x exactly correspond to that in [14]. The modulation is slightly different for pyramidal cells and interneurons.

The results of estimation are presented in Fig.2. It follows that the synapses seem to be more sensitive targets, than membrane ionic channels. Simulation of neural network activity with respect to given radiation dose revealed loss of activity with increase of dose. Typical example is given in Fig.3.

The quantitative characteristic of working memory failure was introduced based on the averaged firing time T_{end} after the stimulation. Thus, unit value of $1 - T_{cue}/T_{end}$ corresponds to infinitely long firing time with respect to stimulus time T_{cue}. The analysis of stability regions presented in Fig.4 shows that the instability arises at the excess of threshold radiation dose. The threshold is more pronounced for heavy LET particles, while for low LET protons the stability region has metastable windows at doses higher than 2Gy. This result correlates with recent experimental findings [23], where uncertainty attending to the

possible disruption of cognitive performance caused by proton irradiation was found. Such effect supports the reliability of our model.

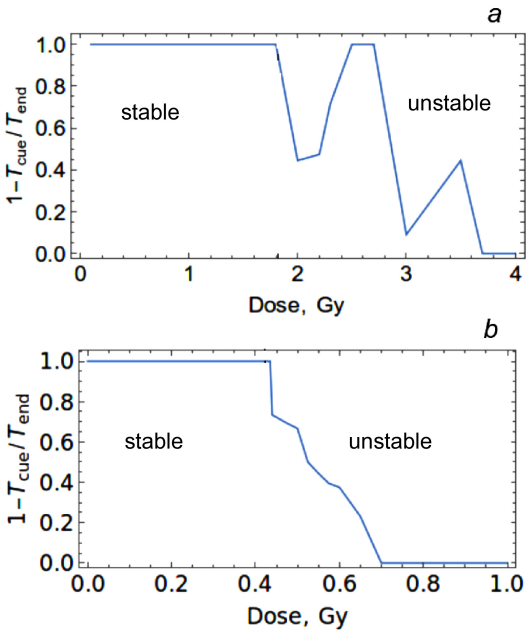


FIG. 4: Estimated working memory stability regions with respect to absorbed dose of 150 MeV/n protons (a) and of 600 MeV/n ^{56}Fe ions (b).

4. Conclusion

We have developed biophysical conductance-based neural network model of working memory for study of radiation-induced impairments. The model can be used to show how different synaptic and voltage-gated conductances contribute to persistent activity, how neuromodulation could influence its robustness, and finally how sustained activity can be stable after the damaging action of external factor such as ionizing radiation.

We have applied phenomenological approach by using interpolated values of dose-dependent changes in basic structural elements of neurons (synaptic receptors, ion channels, etc) according to known experimental data. The simulation of network spatiotemporal dynamics was performed for typical cognitive task. It is demonstrated, that radiation-induced alterations in the properties of synaptic receptors cause loss of stability for specific patterns of activity. This instability arises at the excess of threshold radiation dose, which was shown to be an order higher for heavy charged ions than for protons.

Proposed theoretical approach provides an insight on how can new knowledge and data from molecular, cellular and tissue models of CNS adverse changes be used to estimate CNS risks to astronauts from galactic cosmic rays during the interplanetary flight.

The work of A.N. Bugay was supported by Russian Science Foundation (Project No 17-11-01157).

References

1. Grigoryev, A.I. and Krasavin, E.A. and Ostrovsky, M.A.: On the Evaluation of the Risk of the Biological Action of Galactic Heavy Ions during an Interplanetary Flight. Russ. J. Physiol., Vol. 99, No. 3. 273-280 (2013).
2. Curtis, S.V. and Vazquez, M.E. and Wilson, J.W., et al.: Cosmic Ray Hits in the Central Nervous System at Solar Maximum. Adv. Space Res., Vol. 25. P. 2035–2040 (2000).
3. Greene-Schloesser, D. et al.: Radiation-induced brain injury: a review. Frontiers in oncology, Vol. 2, 1–18 (2012).
4. Shi, L. and Adams, M. M. and Long, A. and Carter, C. C. and Bennett, C. and Sonntag, W. E. and Nicolle, M. M. and Robbins, M. and DAgostino, R. Jr. and Brunso-Bechtold, J.K.: Spatial Learning and Memory Deficits after Whole-Brain Irradiation are Associated with Changes in NMDA Receptor Subunits in the Hippocampus. Radiat. Res. Vol. 166, 892–899 (2006).
5. Machida, M. and Lonart, G. and Britten, R.A.: Low (60 cGy) Doses of ^{56}Fe HZE-Particle Radiation Lead to a Persistent Reduction in the Glutamatergic Readily Releasable Pool in Rat Hippocampal Synaptosomes. Radiat. Res., Vol. 174, 618–623 (2010).
6. Britten, R.A. and Davis, L.K. and Jewell, J.S. and Miller, V.D. and Hadley, M.M. and Sanford, L.D. and Machida, M. and Lonart G.: Exposure to Mission Relevant Doses of 1 GeV/Nucleon ^{56}Fe Particles Leads to Impairment of Attentional Set-Shifting Performance in Socially Mature Rats. Radiation Research, Vol. 182, 292–298 (2014).
7. Dawson, K.B. and Seymour, R. and Mutton, D.E.: Radiation-Induced Efflux of Intracellular Potassium. Radiation Research, Vol.37, 83–89 (1969).
8. Mullin, M.J. and Hunt, W.A. and Harris, R.A.: Ionizing Radiation Alters the Properties of Sodium Channels in Rat Brain Synaptosomes. J. Neurochem., Vol.47, No.2, 489–495 (1986).
9. Hunt, W.A. and Rabin, B.M. and Joseph, J.A. and Dalton, T.K. and Murray, W.E. Jr. and Stevens, S.A.: Effects of Iron Particles on Behavior and Brain Function: Initial Studies. In Terrestrial Space Radiation and its Biological Effects, Plenum Publishing Corporation, (1988).
10. Sokolova, I.V. and Schneider, C.G. and Bezaire, M. and Soltesz, I. and Vlkolinsky, R. and Nelson, G.: Proton radiation alters intrinsic and synaptic properties of CA1 pyramidal neurons of the mouse hippocampus. Radiation Research, Vol. 183, 208-218 (2015).
11. Goldman-Rakic, P.S.: Cellular basis of working memory. Neuron, Vol. 14, 477–485 (1995).
12. Amit, D.J. and Brunel, N.: Model of global spontaneous activity and local structured activity during delay periods in the cerebral cortex. Cereb. Cortex., Vol.7, 237–252 (1997).
13. Brunel, N. and Wang, X.: Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. J. Comput. Neurosci., Vol. 11, 63–85 (2001).
14. Okimura, T. and Tanaka, S. and Maeda, T. and Kato, M. and Mimura, M.: Simulation of the capacity and precision of working memory in the hypodopaminergic state: relevance to schizophrenia. Neuroscience, Vol. 295, 80–89 (2015).

15. Ordy, J.M. and Samorajski T., and Horrocks, L.A. and Zeman, W. and Curtis, H.J.: Changes in memory, electrophysiology, neurochemistry and neuronal ultrastructure after deuteron irradiation of the brain in C57B1-10 mice. J. Neurochem., Vol.15, 1245-1256 (1968) .
16. Amato, A. and Connolly, C.N. and Moss, S.J. and Smart, T.G.: Modulation of Neuronal and Recombinant $GABA_A$ Receptors by Redox Reagents. J Physiol. Vol. 517, 35–50 (1999).
17. Janaky, R and Varga, V. and Saransaari, P. and Oja, S.S.: Glutathione Modulates the N-methyl-D-Aspartate Receptor-Activated Calcium iNflux into Cultured rat Cerebellar Ganule Cells. Neurosci. Lett. Vol. 156, 153–157 (1993).
18. Rabin, B.M., and Joseph, J.A. and Shukitt-Hale, B. and McEwen, J.: Effects of exposure to heavy particles on a behavior mediated by the dopaminergic system. Adv. Space. Res. Vol. 25, 2065–2074 (2000).
19. Shtemberg, A.S. and Kokhan, V.S. and Kudrin, V.S. and Matveeva, M.I and Lebedeva-Georgievskaya, K.D. and Timoshenko, G.N. and Molokanov, A.G. and Krasavin, E.A. and Narkevich, V.B. and Klodt, P.M. and Bazyan, A.S: The Effect of High Energy Protons in the Bragg Peak on the Behavior of Rats and the Exchange of Monoamines in Some Brain Structures. Neurochemical Journal, Vol. 9, No. 1, 66–72 (2015).
20. Batmunkh, M. and Belov, O.V. and Bayarchimeg, L. and Lhagva, O. and Sweilam, N.H.: Estimation of the spatial energy deposition in CA1 pyramidal neurons under exposure to ^{12}C and ^{56}Fe ion beams. J.Radiat. Res. Appl. Sci. Vol.8, 498–507 (2015).
21. Alp, M. and Parihar, V.K. and Limoli, C.L. and Cucinotta, F.A.: Irradiation of Neurons with High-Energy Charged Particles: An In Silico Modeling Approach. PLoS Comput. Biol. Vol.11, e1004428 (2015)
22. Limoli, C.L. and Giedzinski, E. and Baure, J. and Rola, R. and Fike, J.R.: Redox changes induced in hippocampal precursor cells by heavy ion irradiation. Radiat. Environ. Biophys. Vol.46, No.2, 167-172 (2007).
23. Rabin, B.M., and Heroux, N.A. and Shukitt-Hale, B. and Carrihill-Knoll, K.L. and Beck, Z. and Baxter, C.: Lack of reliability in the disruption of cognitive performance following exposure to protons. Radiat Environ. Biophys. Vol. 54, No.3, 285–295 (2015).

On Similarity related to the Genetic Code

Branko Dragovich[1,2] and Nataša Ž. Mišić[3]

[1] Institute of Physics, University of Belgrade, Belgrade, Serbia
[2] Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
dragovich@ipb.ac.rs
[3] Research and Development Institute Lola Ltd, Belgrade, Serbia
nmisic@rcub.bg.ac.rs

Abstract. We consider a few kinds of distance for description of (dis)similarity in the genetic code. We point out relevance of ultrametrics, and especially p-adic distance, for modeling the genetic code and investigation of similarity between sequences of nucleotides, codons or amino acids.

Keywords: p-genetic code, ultrametric similarity, p-adic similarity.

1. Introduction

In living organisms, particularly in bioinformation, there is often some relation between structure and function. Mainly similar structure implies similar function. How to measure similarity? Two things which look similar are in some sense close (near) each other. Closeness is usually measured by a distance. In the case of biological similarity a Euclidian metric is not appropriate. We shall here consider some metrics which are more or less suitable to measure similarity.

In this article we are interested in similarity between some biomolecular sequences. These sequences can be composed either of nucleotides, codons or amino acids.

In Sec. 2 we consider distances which are somehow relevant for measuring (dis)similarity between sequences and we present some their general properties. Similarity between sequences of nucleotides which make codons is subject of Sec. 3. Similarities between sequences of nucleotides and codons will be considered in Sec. 4. At the end of this article are some concluding remarks.

2. Distances for (Dis)similarities

Let M be a set of some elements denoted by $x, y, z, \ldots$ Recall that distance d is a real-valued function defined for any two elements of M, which satisfies the following properties:

$$(i)\ d(x,y) \geq 0, \quad d(x,y) = 0 \Leftrightarrow x = y, \tag{1}$$

$$(ii)\ d(x,y) = d(y,x), \tag{2}$$

$$(iii)\ d(x,y) \leq d(x,z) + d(z,y). \tag{3}$$

The last property is called triangle inequality. Metric space is a pair (M,d).

How to quantify similarity by a distance? Looking along two sequences of the equal length, we intuitively understand that their similarity depends on the number of positions with the same elements: more such positions – more similarity. This property of two sequences, we shall call *sequence similarity*. In the case of the genetic code, we shall see that *functional similarity* depends not only on the number of such positions but also on the place in codons where it happens, and that the beginning of codons is more important than their end. Dissimilarity is an opposite property.

Let $W_{k,n}(N)$ be a set of sequences (words) of equal length n composed of k different elements (letters). Then total number of elements (words) of this set (language) is $N = k^n$. Denote elements of $W_{k,n}(N)$ in the form $x = x_1 x_2 \cdots x_n$.

The Hamming distance, introduced in 1950, between two elements $a = a_1 a_2 \cdots a_n$ and $b = b_1 b_2 \cdots b_n$ 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 $a_i \neq b_i$. That is $d_H(a,b) = n - \nu$, where ν is the number of positions at which elements of both sequences are equal. So, smaller distance – closer sequences. In information theory, this distance corresponds to the minimum number of substitutions necessary to change one sequence into the other. Pair $(W_{k,n}(N), d_H)$ is a metric space.

The Levenshtein distance, introduced in 1965, is the minimum number of insertions, deletions and substitutions required to transform one sequence into the other: $d_L(a,b) = \min(n_I + n_D + n_S)$, where n_I, n_D, n_S denote the number of insertions, deletions and substitutions, respectively. In the case of Levenshtein distance it is not necessary that sequences have the same length. When two sequences have the same length then the Levenshtein distance is smaller or equal to the Hamming distance.

The Damerau-Levenshtein distance is the Levenshtein distance extended by inclusion of transpositions of two adjacent elements in one of sequences, i.e. $d_{DL}(a,b) = \min(n_I + n_D + n_S + n_T)$, where n_T denotes the number of transpositions.

The above metrics are examples of edit distances which measure the minimum number of operations required to transform one sequence (word) into the other.

2.1. Ultrametric distance

Now we want to present a subclass of metrics spaces, which is called ultrametric space, and demonstrate that p-adic distance is more adequate for characterization of bioinformation similarity than the above mentioned distances.

An ultrametric space is a metric space if its distance also satisfies ultrametric (strong triangle, non-Archimedean) inequality

$$d(x,y) \leq \max\{d(x,z), d(z,y)\}. \tag{4}$$

Ultrametric space was introduced in 1944 by M. Krasner (1912–1985). Note that some aspects of ultrametric spaces have been used earlier. For example, taxonomy contains ultrametrics and it started 1735 by C. Linné's (1707–1778) biological classification with hierarchical structure. Namely, living organisms with more common ancestors are ultrametrically closer than those with less ones.

As a consequence of the ultrametric inequality (4), ultrametric spaces have many unusual properties. For example, by suitable notation of points x, y, z, inequality (4) can be rewritten in the form $d(x,y) \leq d(x,z) = d(y,z)$. This means that all ultrametric triangles are isosceles. Recall also the following properties: (i) There is no partial intersection of the balls; (ii) Any point of a ball can be treated as its center; (iii) Each ball is both open and closed (clopen). For a proof of these properties of ultrametric balls, see e.g. Schikhof's book [1].

Let us illustrate ultrametric distance using the above introduced set of words $W_{k,n}(k^n)$. In particular, we take that number of letters $k = 4$ and number of letters in words is $n = 3$, i.e. we take the case $W_{4,3}(64)$. We briefly consider three examples of ultrametric distance: ordinary ultrametric distance, the Baire distance and p-adic distance.

Ordinary ultrametric distance. Ordinary ultrametric distance between any two different words x and y is $d(x,y) = n-(m-1)$, where $m (m = 1,2,...,n)$ is the first position at which letters differ counting from the beginning. It has n values, i.e. $d(x,y) = 1,2,...,n$. This distance can be scaled as $d_s(x,y) = \frac{n-m+1}{n}$ and then it takes values: $1, \frac{n-1}{n}, ..., \frac{2}{n}, \frac{1}{n}$.

In the particular case $W_{4,3}(64)$ there are 64 three-letter words. Let the four letters be a,b,c,d (see Table 1). Possible distances between words x and y are $d(x,y) = 1,2,3$ and the corresponding scaling ones are: $d_s(x,y) = 1, \frac{2}{3}, \frac{1}{3}$. For example, $d_s(abc,bac) = 1$, $d_s(abc,acb) = \frac{2}{3}$, $d_s(abc,abb) = \frac{1}{3}$.

The Baire distance. This distance is usually defined as $d_B(x,y) = 2^{-(m-1)}$, where m is the first position in words x and y at which letters differ, i.e. $m = 1,2,...,n$. Thus the Baire distance takes values: $1, \frac{1}{2}, \frac{1}{2^2}, ..., \frac{1}{2^{n-1}}$. Instead of the base 2 one can take any natural number larger than 2.

In the case $W_{4,3}(64)$ the Baire distance has values: $1, \frac{1}{2}, \frac{1}{4}$. For example, $d_B(abc,bab) = 1$, $d_B(abc,acb) = \frac{1}{2}$, $d_B(abc,abb) = \frac{1}{4}$.

2.2. p-Adic distance

The most important class of ultrametric spaces contains fields $\mathbb{Q}_p$ of p-adic numbers which were introduced in 1897 by K. Hensel (1861–1941).

By definition, p-adic absolute value (p-adic norm) of a non-zero integer $u \in \mathbb{Z}$ is $|u|_p = p^{-k}$, where k is degree of a prime number p in u, and $|0|_p = 0$. Since $k = 0,1,2,...$, p-adic absolute value of any integer u is $|u|_p \leq 1$ and this is valid for any prime number p. p-Adic distance between two integers u and v is $d_p(u,v) = |u-v|_p$. This distance is related to divisibility of $u-v$ by prime p (more divisible – lesser distance). With respect to a fixed prime p as a base, any natural number has its unique expansion $u = u_0 + u_1 p + u_2 p^2 + ... + u_n p^n$, where $u_i \in \{0,1,...,p-1\}$ are digits. If in this expansion k is the smallest degree, then p-adic norm of u is $|u|_p = p^{-k}$. Note that p-adic distance, as any other ultrametric distance, has discrete values.

The above three-letter words can be connected with three-digit numbers by identifying letters a,b,c,d with four digits $1,2,3,4$ in the following way: $a =$

$1, b = 2, c = 3, d = 4$. Number $p = 5$ is the smallest prime number, which taken as an expansion base contains four digits $\{1, 2, 3, 4\}$ and digit 0, which can be ignored (see Table 1).

One can construct a set of 5-adic integers in the form

$$x = x_0 + x_1 5 + ... + x_k 5^k \quad \text{or} \quad x \equiv x_0 x_1 ... x_k, \quad x_i \in \{1,2,3,4\}. \tag{5}$$

In (5) positional notation (encoding) of natural numbers is opposite to the usual one. Note that the four letters $\{a,b,c,d\}$ can be identified with the four digits $\{1,2,3,4\}$ in 24 different ways. We use: $a = 1, b = 2, c = 3, d = 4$. Then there are 64 words presented in two different ways – by three letters and three digits, see Table 1.

In the case $W_{4,3}(64)$ there are three-letter words represented now by three-digit 5-adic numbers (Table 1). The corresponding 5-adic distance of a pair of words (numbers) $x = x_0x_1x_2 \equiv x_0 + x_1 5 + x_2 5^2$ and $y = y_0y_1y_2 \equiv y_0 + y_1 5 + y_2 5^2$ is

$$d_5(x,y) = |x_0x_1x_2 - y_0y_1y_2)|_5 = \begin{cases} 1, & x_0 \neq y_0 \\ \frac{1}{5}, & x_0 = y_0, x_1 \neq y_1 \\ \frac{1}{25}, & x_0 = y_0, x_1 = y_1, x_2 \neq y_2 \,. \end{cases} \tag{6}$$

For example, $d_5(123,213) = 1$, $d_5(123,132) = \frac{1}{5}$, $d_5(123,122) = \frac{1}{25}$.

As we shall see later, p-adic distance between sequences of biomolecules is finer and more informative than the ordinary ultrametric and the Baire distance. Namely, for the same set of natural numbers one can also employ p-adic distance with $p \neq 5$, in particular we use $p = 2$.

It is worth mentioning that the most advanced examples of the ultrametric spaces are the fields of p-adic numbers $\mathbb{Q}_p$, where index p denotes any prime number. There are infinitely many fields $\mathbb{Q}_p$ which are not mutually isomorphic – for every prime number p there is its own $\mathbb{Q}_p$. The field $\mathbb{Q}_p$ can be constructed by completion of the field $\mathbb{Q}$ of rational numbers in the same way as it is usually done for the field $\mathbb{R}$ of real numbers, just one has to take $|\cdot|_p$ instead of the usual absolute value $|\cdot|$. p-Adic numbers and their functions are rather well developed part of modern mathematics, see e.g. books [2, 1]. Many applications from Planck scale physics via complex systems to the universe as a whole, known as p-adic mathematical physics, have been considered, e.g. see recent review articles [3, 4]. p-Adic and standard models (over real and complex numbers) are connected within adelic framework, see adelic quantum mechanics [5, 6]. In this article devoted to similarity within the genetic code, and an extension to similarity between some sequences of biomolecules, only p-adic distance is used. For p-adic modeling of the genetic code see [7–11] and [12].

The above examples illustrate how ultrametric distance measures (dis)similarity between two words, i.e. (dis)similarity between two elements of an ultrametric space. Also these ultrametric examples can be represented by trees. Namely, instead of the four letters $\{a,b,c,d\}$ or digits $\{1,2,3,4\}$ in the three-letter words one can take line segments to draw four edges of the related tree (see Fig. 1).

TABLE 1: Table of three-letter words constructed of four letters and arranged in the ultrametric form. The same has done for the corresponding three-digit 5-adic numbers, where four digits are identified as $a = 1, b = 2, c = 3, d = 4$. These are two representations of ultrametric space $W_{4,3}(64)$. Here 64 three-digit 5-adic numbers (three-letter words) are presented so that within quadruplets 5-adic distance is the smallest, i.e. $d_5(x,y) = \frac{1}{25}$, while 5-adic distance between any two quadruplets in vertical line is $\frac{1}{5}$ and otherwise is equal 1. Ultrametric tree illustration of these cases is in Fig. 1.

111 **aaa**	211 **baa**	311 **caa**	411 **daa**
112 **aab**	212 **bab**	312 **cab**	412 **dab**
113 **aac**	213 **bac**	313 **cac**	413 **dac**
114 **aad**	214 **bad**	314 **cad**	414 **dad**
121 **aba**	221 **bba**	321 **cba**	421 **dba**
122 **abb**	222 **bbb**	322 **cbb**	422 **dbb**
123 **abc**	223 **bbc**	323 **cbc**	423 **dbc**
124 **abd**	224 **bbd**	324 **cbd**	424 **dbd**
131 **aca**	231 **bca**	331 **cca**	431 **dca**
132 **acb**	232 **bcb**	332 **ccb**	432 **dcb**
133 **acc**	233 **bcc**	333 **ccc**	433 **dcc**
134 **acd**	234 **bcd**	334 **ccd**	434 **dcd**
141 **ada**	241 **bda**	341 **cda**	441 **dda**
142 **adb**	242 **bdb**	342 **cdb**	442 **ddb**
143 **adc**	243 **bdc**	343 **cdc**	443 **ddc**
144 **add**	244 **bdd**	344 **cdd**	444 **ddd**

3. Similarities within the Genetic Code

Now, we want to apply p-adic ultrametric space $W_{4,3}(64)$ to modeling the genetic code. Before to do that, it is useful to recall a few basic properties of the genetic code, particularly of the vertebrate mitochondrial code.

The genetic code is a rule which tells us how 20 amino acids (building blocks of proteins) and one stop signal are coded by 64 codons (building blocks of genes). From mathematical point of view, the genetic code is a map from a set of 64 elements onto a set of 21 element. According to an estimation, there is about 1.5×10^{84} possibilities for genetic coding, while living organisms use practically one code with few dozen slight variations. The aim of the genetic code modeling is to find an adequate mathematical description of the codes in living organisms.

Codons are ordered triplets made of four nucleotides C (Cytosine), A (Adenine), G (Guanine), and T (Thymine) or U (Uracil). There are $4 \times 4 \times 4 = 64$ codons, and each of them codes an amino acid or stop signal in the process of the protein synthesis in ribosomes.

In human cells there are two codes: standard and vertebrate mitochondrial (VM) code. The VM code is simpler than the standard one and all other codes

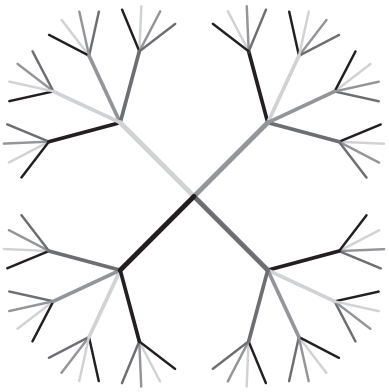


FIG. 1: Utrametric tree related to Table 1 with $W_{4,3}(64)$ case and also to the vertebrate mitochondrial code presented at the Table 2. One can easily calculate ordinary ultrametric distance and see that distance between any three end points satisfies the strong triangle (ultrametric) inequality.

can be regarded as slight variations of this one. In the VM code 64 codons are arranged into 32 doublets. These doublets have the same nucleotides at the first two positions, while at the third position is a nucleotide, which is purine or pyrimidine. Hence, there are 16 doublets with purine and 16 doublets with pyrimidine at the third position. Each of these codon doublets codes an amino acid or stop signal. Since there are 32 doublets and 20 amino acids with one stop signal, it means that some amino acids (or stop signal) are coded by more that one codon doublet. In the VM code there are: 12 amino acids coded by single doublets, 6 amino acids and stop signal coded by two doublets, and 2 amino acids coded by three doublets (see Table 2). This property that some amino acids are coded by more than one codon is called *code degeneracy* and it is very important for optimization of the genetic code for robustness to translation errors and mutations.

Since in a doublet codons code the same amino acid or stop signal, it means that these codons have the same coding function. In fact, this equality in functioning is a result of similarity in codon's sequence structure. Namely, as it is already said, codons in a doublet have the same nucleotides at the first two positions, and a small difference in nucleotides at the third position. To describe quantitatively this similarity we use p-adic distance with both $p = 5$ and $p = 2$. It is also important to find appropriate identification of nucleotides $C, A, T(U), G$ with digits $1, 2, 3, 4$ in 5-adic expansion of 64 natural numbers in the form (5) when $k = 3$. We use the following identification: $C = 1, A = 2, T = U = 3, G = 4$. Codons can be also connected with ultrametric space of words $W_{4,3}(64)$ taking $a = C, b = A, c = T = U, d = G$.

Now we can see that codon space of the VM code has the same ultrametric structure as the set of words $W_{4,3}(64)$, cf. Tab. 1 and Tab. 2. According to the smallest 5-adic distance, which is $\frac{1}{25}$, one obtains quadruplets of codons. Then

TABLE 2: Here is the vertebrate mitochondrial code with p-adic ultrametric structure. Digits are related to nucleotides as follows: $C = 1, A = 2, U = 3, G = 4$. 5-Adic distance between codons: $\frac{1}{25}$ inside quadruplets, $\frac{1}{5}$ between different quadruplets in the same column, 1 otherwise. Each quadruplet can be viewed as two doublets, where every doublet codes an amino acid or stop (termination) signal (Ter). 2-Adic distance between codons in doublets is $\frac{1}{2}$. Amino acids which are coded by two doublets are in fact coded by the corresponding quadruplet. Amino acids leucine (Leu) and serine (Ser) are coded by three doublets – the third doublet is at $\frac{1}{2}$ 2-adic distance with respect to the corresponding doublet in quadruplet.

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

application of 2-adic distance inside quadruplets results in separation of any quadruplet to two corresponding doublets. Codons inside any doublet are at $\frac{1}{2}$ 2-adic distance, which is the smallest 2-adic distance inside quadruplets. Thus, combination of 5-adic and 2-adic distance between codons is a simple and adequate mathematical tool to express their structural similarity relevant to their functional similarity.

4. Similarities between sequences of nucleotides or codons

Now one can ask question: How to apply some of the above distances to investigation of similarity between any two sequences of nucleotides or sequences of codons, with the same length? This can be considered as investigation of similarity between words of a set $W_{k,n}(N)$, where elements (letters) of words are nucleotides or codons.

4.1. Ultrametric approach

Let words of $W_{k,n}(N)$ are sequences which elements are nucleotides or codons. Then ordinary ultrametric distance between two different sequences $x = x_1x_2...x_n$ and $y = y_1y_2...y_n$ may have one of the following n values:

$$d(x,y) = d(x_1x_2...x_n, y_1y_2...y_n) = 1, 2, 3, ..., n, \tag{7}$$

where $d(x,y) = n-k$ $(0 \leq k \leq n-1)$ if $x_i = y_i$ for all indices $i < k+1$ and $x_{k+1} \neq y_{k+1}$.

In the case that elements of sequences are nucleotides one can extend the above approach with 5-adic distance, which may have one of the following n values:

$$d_5(x,y) = d_5(x_0x_1...x_{n-1}, y_0y_1...y_{n-1}) = 1, 5^{-1}, 5^{-2}, ..., 5^{-(n-1)}. \tag{8}$$

One can also use the Baire distance (see definition in Sec. 2).

4.2. Modified Hamming distance

The Hamming distance is defined in Sec. 2. Under modified Hamming distance we understand Hamming like distance where $d(x_i, y_i) = 1$ can be replaced by $d(x_i, y_i) < 1$ when it makes sense.

For example, in the case of two sequences of nucleotides when $x_i = C = 1$ and $y_i = U = 3$, or $x_i = A = 2$ and $y_i = G = 4$, then there is a sense to take $d_2(C,U) = |3-1|_2 = |2|_2 = \frac{1}{2}$ and analogously for $d_2(A,G) = |4-2|_2 = \frac{1}{2}$. Namely, nucleotides C and U are chemically more similar than e.g. C or U and A or G. C and U are pyrimidines, while A and G are purines.

If elements of sequences are codons instead of nucleotides then there is a sense to take distances between codons as their 5-adic distances rather than usual Hamming distance. Moreover, there are codons which code the same amino acid and also some codons that code different amino acids, as it is pointed out in the previous section.

In this way modified Hamming distance is finer than its standard form.

5. Concluding Remarks

In this article we have presented a few metrics, more or less appropriate to characterize similarity between sequences, which elements may be nucleotides or codons. The above consideration can be extended to investigation of similarity between sequences of amino acids.

Note that the Hamming distance is not appropriate to characterize similarity between codons with respect to their information content, because it does not take into account place of positions at which two nucleotides are equal or different.

It seems to be useful to introduce measure of similarity $\mathscr{S}$ so that it has values $0 \leq \mathscr{S} \leq 1$. If so, then on metric space $W_{k,n}(N)$ similarity between given sequences x and y can be defined as $\mathscr{S}(x,y) = \frac{d_{max} - d(x,y)}{d_{max}}$, where d_{max} is maximal distance at the space $W_{k,n}(N)$. It is also natural to introduce dissimilarity $\bar{\mathscr{S}}(x,y)$ as $\bar{\mathscr{S}}(x,y) = \frac{d(x,y)}{d_{max}}$. It holds $\mathscr{S}(x,y) + \bar{\mathscr{S}}(x,y) = 1$.

Acknowledgments

This work was supported in part by Ministry of Education, Science and Technological Development of the Republic of Serbia, projects: OI 173052, OI 174012, TR 32040 and TR 35023.

References

1. W. H. Schikhof, *Ultrametric Calculus: An Introduction to p-Adic Calculus* (Cambridge University Press, 1984).
2. I. M. Gel'fand, M. I. Graev and I. I. Pyatetski-Shapiro, *Representation Theory and Automorphic Functions* (Saunders, Philadelphia, 1969).
3. B. Dragovich, A. Yu. Khrennikov, S. V. Kozyrev and I. V. Volovich, "On *p*-adic mathematical physics", *p-Adic Numbers Ultrametric Anal. Appl.* **1** (1), 1–17 (2009), arXiv:0904.4205v1[math-ph].
4. B. Dragovich, A. Yu. Khrennikov, S. V. Kozyrev, I. V. Volovich and E. I. Zelenov, "*p*-Adic mathematical physics: the first 30 years", *p-Adic Numbers Ultrametric Anal. Appl.* **9** (2), 87–121 (2017), arXiv:1705.04758[math-ph].
5. B. Dragovich, "Adelic model of harmonic oscillator", *Theor. Math. Phys.* **101**, 1404–1415 (1994), arXiv:hep-th/0402193.
6. B. Dragovich, "Adelic harmonic oscillator", *Int. J. Mod. Phys. A* **10**, 2349–2365 (1995), arXiv:hep-th/0404160.
7. B. Dragovich and A. Dragovich, "A *p*-adic model of DNA sequence and genetic code", *p-Adic Numbers Ultrametric Anal. Appl.* **1** (1), 34–41 (2009), arXiv:q-bio.GN/0607018v1.
8. B. Dragovich and A. Dragovich, "*p*-Adic modelling of the genome and the genetic code", *Computer J.* **53** (4), 432–442 (2010), arXiv:0707.3043v1 [q-bio.OT].
9. B. Dragovich, "*p*-Adic structure of the genetic code", (2012), arXiv:1202.2353 [q-bio.OT].
10. B. Dragovich, "Genetic code and number theory", *Facta Universitatis: Phys. Chem. Techn.* **14** (3), 225–241, (2016), arXiv:0911.4014 [q-bio.OT].
11. B. Dragovich, A. Yu. Khrennokov and N. Ž. Mišić, "Ultrametrics in the genetic code and the genome", *Appl. Math. Comput.* **309**, 350–358 (2017), arXiv:1704.04194 [q-bio.OT].
12. A. Khrennikov and S. Kozyrev, "Genetic code on a diadic plane", *Physica A: Stat. Mech. Appl.* **381**, 265–272 (2007), arXiv:q-bio/0701007.

Machine learning-based approach to help diagnosing Alzheimer's disease through spontaneous speech analysis

Jelena Graovac[1], Jovana Kovačević[1], and Gordana Pavlović Lažetić[1]

Faculty of Mathematics, University of Belgrade, Studentski trg 16
11000 Belgrade, Serbia
{jgraovac,jovana,gordana}@matf.bg.ac.rs

Abstract

Alzheimer's disease and other dementias have been recognized as a major public health problem among the elderly in developing countries. We address this issue by exploring automatic noninvasive techniques for diagnosing patients through analysis of spontaneous, conversational speech. The technique we are proposing is a variant of n-gram based kNN machine learning technique. Since we use byte-level n-grams, we do not use any language dependent information, including word boundaries, character case, white-space characters or punctuation.

Twelve adults diagnosed with dementia of Alzheimer type (DAT) participate in the study. All DAT participants were interviewed at adult day care center for people with Alzheimer's disease or dementia in Novi Sad, the only institution of its kind in Serbia. All interviews were audio-recorded, transcribed verbatim by a trained researcher, and checked for accuracy by the authors. Means for the Mini-Mental Status Exam distinguished the two groups: moderate and mild.

Our plan is to compile a control dataset based on the interviews of healthy elderly that do not differ significantly in age, sex or education level from the DAT participants. We plan to compare DAT and healthy elderly participants to test how well our techniques will discriminate between these groups. In this paper, we make some preliminary distinction between the two groups of the DAT participants. Our plan is to develop new, more sophisticated classification techniques, based on Machine Learning and Natural Language Processing. We hope that our techniques will show promising as diagnostic and prognostic additional tools that may help earlier diagnosis of DAT and determining its degree of severity.

Keywords: dementia of Alzheimer type, automatic diagnostics, natural language processing, machine learning

1. Introduction

According to the last census (2015) Serbia has a population of 7.1 million people, of which 19.7% are over 65 years old. It is estimated that around 200,000

people live with dementia, although epidemiological studies have not been conducted yet. Many studies have shown that there is a need for more reliable diagnosis, as well as the education of professionals and the public. A significant component of the dementia of Alzheimer type (DAT) that accompanies Alzheimer's disease is aphasia, a loss of oral and written communicative ability. Symptoms include shallow vocabularies and word-finding difficulties leading to the deterioration of spontaneous speech which is often observed by family members during conversational situations in the early stage of disease [5]. Lately, there are several encouraging reports about applying machine learning algorithms for automatic diagnosis of DAT based on spontaneous speech analysis [4], [1], [3].

We address this issue by exploring automatic machine learning noninvasive method for diagnosing patients through analysis of spontaneous, conversational speech. The technique we are proposing here belongs to group of n-gram based kNN techniques. Using byte-level n-grams, we are enabled to avoid any language dependent information, including word boundaries, character case, white-space characters or punctuation. However, white-space and punctuation characters implicitly play a significant role in classier performance based on the frequency of occurrence [5].

2. Methodology and data

2.1. Dataset

Twelve adults diagnosed with DAT (in the Clinical Center of Vojvodina, Novi Sad) participate in the study. All DAT participants were interviewed at the adult day care center for people with Alzheimer's disease or dementia in Novi Sad, the only institution of its kind in Serbia. All interviews were audio-taped, transcribed verbatim by a trained researcher, and checked for accuracy by the authors. Means for the Mini-Mental Status Exam (MMSE) distinguished the two groups: moderate (score between 10 and 18) and mild (score between 19 and 23). The MMSE involves a patient responding to 17 questions that cover a wide range of cognitive domains divided into two sections: the first requires verbal responses to orientation, memory, and attention questions, and second section requires reading and writing [5]. All interviews are saved in the separate documents and divided into test set and training set.

Our plan is to compile a control dataset based on the interviews of healthy elderly that do not differ significantly in age, sex or education level from the DAT participants.

2.2. Our technique

In this research we used ngram-based technique, presented and used by Keselj et al. [2] to solve the authorship attribution problem. The technique is based on byte-level n-gram frequency statistics method for document representation, and kNN ($k = 1$) algorithm for text classification process. Extracting byte n-grams from a document is like moving an n-byte wide "window" across the document,

byte by byte. Each window position covers n bytes, defining a single n-gram. N-gram techniques have been successfully used for a long time in a wide variety of problems and domains.

Preprocessing All texts from the training set are concatenated in one of two meta-texts depending on a category. For each text from the test set, each text from the training set and for the two meta-texts, a profile representation are generated (meta-text profiles are named category profiles). Each profile consists of first L most frequent byte n-grams encoded in a pair (x_i, f_i), where x_i is one byte n-gram and f_i its normalized frequency. The profile length L limits the number of n-grams considered during the similarity calculation and serves to keep profiles small when large values of n are used.

Training phase For different values of n and L, for each train document:

- Compute a dissimilarity measure between the train documents profile and each of the categorys profiles.
- Select the category (or categories) whose profile has the smallest value of dissimilarity measure with the documents profile
- Make evaluation of the obtained results
- Choose n and L with the best obtained results.

Test phase For obtained n and L, for each test document:

- Compute a dissimilarity measure between the test documents profile and each of the categorys profiles.
- Select the category (or categories) whose profile has the smallest value of dissimilarity measure with the documents profile.

In order to decide whether a certain test document belongs (or not) to a certain category, this text classification procedure requires a dissimilarity measure. In this paper we used measure presented by Keselj et al. [2] that has a form of relative distance:

$$d(\mathscr{P}_1, \mathscr{P}_2) = \sum_{n \in profile} \left(\frac{2 \cdot (f_1(n) - f_2(n))}{f_1(n) + f_2(n)} \right)^2 \tag{1}$$

where $f_1(n)$ and $f_2(n)$ are frequencies of an n-gram n in the category profile $\mathscr{P}_1$ and the test document profile $\mathscr{P}_2$, respectively.

2.3. Performance evaluation

For evaluating the performance of the technique, the typical evaluation metrics that come from information retrieval are used: P recision (P), Recall (R) and F1 measure:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F1 = \frac{2PR}{P+R} \tag{2}$$

where TP (True Positives) are defined as the documents that were correctly assigned to the considered category while FP (False Positives) are the documents that were wrongly assigned to that category. Similarly, TN (True Negatives) were correctly not assigned to the considered category, while FN (False Negatives) were not assigned to the considered category but should have been assigned to it (since they belong to it). All presented measures can be aggregated over all categories in two ways: micro-averaging the global calculation of measure considering all the documents as a single dataset regardless of categories, and macro-averaging the average on measure scores of all the categories. In this article, micro-averaged F1 and macro-averaged F1 measures are reported.

3. Results

Since we have data only for DAT participants, we are able to test our technique only in making distinction between the two groups of the DAT participants (moderate and mild). Experiments are conducted for different values of n nad L (n take values from interval [2,9], while L take values from interval [100, 50000] with step 100). We used 10-cross validation technique and we obtained following results: micro-averaged F1 = 70% and macro-averaged F1 = 64.66%. Optimal values for parameter n lying between 3 and 6, while for parameter L lying between 1000 and 2000. Authors in [5] obtained accuracy between 53.8% and 69.6% in rating dementia in two classes, so our results are comparable to results presented in this paper.

4. Conclusion and future work

Our goal is to compare DAT and healthy elderly participants to test how well our method will discriminate between these groups. So, our first task is to collect interviews with healthy elderly participants. Moreover, we will work on getting new data about DAT participants. We also plan to develop and test new Machine Learning classification techniques and Natural Language Processing techniques.

Although more work needs to be done to collect new data and improve the accuracy of presented results, we hope that our techniques will show promising as diagnostic and prognostic additional tools that may help earlier diagnosis of DAT and determining its degree of severity.

Acknowledgments

This work was supported in part by Ministry of Education, Science and Technological Development of the Republic of Serbia, projects No. 174021, III44006 and III47003.

References

1. Fraser, Kathleen C and Meltzer, Jed A and Rudzicz, Frank: Linguistic Features Identify Alzheimer's Disease in Narrative Speech. Journal of Alzheimer's Disease, IOS Press, Vol. 49, No. 2, pages 407-422 (2015)

2. Keselj, V., Peng, F., Cercone, N., and Thomas, C. N-gram-based Author Profiles for Authorship Attribution. In Proceedings of the Conference on Pacific Association for Computational Linguistics, PACLING. Halifax, Canada, pp. 255264 (2003).
3. Lopez-de-Ipi na, Karmele et al.: On automatic diagnosis of Alzheimer's disease based on spontaneous speech analysis and emotional temperature. Cognitive Computation, Springer, Vol. 7, No. 1, pages 44-55 (2015)
4. Nasrolahzadeh, Mahda and Mohammadpoori, Zeinab and Haddadnia, Javad: Analysis of mean square error surface and its corresponding contour plots of spontaneous speech signals in Alzheimer's disease with adaptive wiener lter. Computers in Human Behavior, Elsevier, Vol. 61, pages 364-371 (2016)
5. Thomas, Calvin et al.: Automatic detection and rating of dementia of Alzheimer type through lexical analysis of spontaneous speech. Mechatronics and Automation, 2005 IEEE International Conference. Vol. 3. IEEE (2005)

Improving 1NN strategy for classification of some prokaryotic organisms

M. Grbić[1], A. Kartelj[2], D. Matić[1], and V. Filipović[2]

[1] Faculty of Science and Mathematicas, University of Banja Luka, Mladena Stojanovića 2, 78000 Banja Luka, Republic of Srpska, Bosnia and Herzegovina
milanagrbic@yahoo.com, matic.dragan@gmail.com
[2] Faculty of Mathematics, University of Belgrade, Studentski trg 16 11000 Belgrade, Serbia
{kartelj, vladaf}@matf.bg.ac.rs

Abstract. Classification algorithms are intensively used in discovering new information in large sets of biological data. In cases when classification tasks involve nominal attributes, some of commonly used classification tools do not obtain results of satisfying quality, since mathematical operations and relations can not be directly applied to symbolic values. This problem often appears in the k-nearest neighborhood (KNN) classification because the standard Euclidean distance function can become burdened by the large number of irrelevant attributes, consequently producing inaccurate classification results.
In this paper we examine several metrics which can be applied to nominal attributes and for each metric we apply the appropriate KNN strategy. In order to justify the proposed approach, comprehensive experiments are performed on a dataset of prokaryiotic organisms. Experimental results indicate that the new classifications are more accurate than those obtained by the previously used methods, getting better results in seven of total of twelve cases.

Keywords: bioinformatics, classification, nearest neighbor, distance metrics, data mining

1. Introduction

There is a fast growth in the volume of data stored in biological databases. One of the particularly active area in bioinformatics is the development and application of the machine learning methods and classification algorithms in order to obtain more useful information from large sets of biological data.

During the classification process, the classifier uses a set of training records with known classes in order to learn how to predict the class of an record with an unknown class. During past decades, many power and robust classification tools have appeared on the market. Some of the most popular and frequently used such tools offers various techniques, allowing users to try and compare different machine learning methods on new and existing data sets.

Among many other commonly used software framework used for classification, we notice some of them: WEKA [1], KNIME [2], IBM SPSS [3] and IBM Intelligent Miner package [4].

Although these tools are proved to be reliable in many tasks, in classifications which involve many nominal attributes, these tools often do not obtain results of satisfying quality, since mathematical operations and relations need additional customization with respect to the nature of the considered data. As a consequence, such classifiers ignore nominal attributes and form the classification model based solely on numerical attributes. This approach usually leads to inaccurate and unreliable results. The problem of inability to handle nominal attributes especially appears in the classification algorithms based on the KNN strategy. KNN involves a distance function that measures the difference or similarity between two records. In KKN, there is an assumption that the class of a test record is equal to the most frequent class of the nearby records with respect to distance function, e.g. Euclidean distance function. When the classification algorithm has to deal with many nominal attributes, the calculation of the distance between two records can become burdened by the large number of irrelevant attributes. In such cases, we get inaccurate classification results or even no result at all, if all attributes are nominal, since the application of KNN strategy is impossible in such case.

To overcome these problems and enable the application of a KNN classifier to such datasets, new distance functions between attributes needs to be defined. In this paper we examine several metrics known in the literature, which can be applied to nominal attributes of a dataset of prokaryotic organisms.

The dataset analyzed in this paper consists of prokaryotic organisms and contains total of 30 attributes, from which 11 attributes are nominal. Earlier experiments presented in [5, 6] indicated that commonly used classification tools, mostly ignore nominal attributes and forms the classification based on only numerical ones. For each analyzed metric we apply the appropriate KNN strategy, enabling the classification process become more accurate.

This paper is organized as follows. In the next section we present a short description of the KNN method, as well as the overview of the metrics which are convenient for use in determining the distance between the considered data. In the section Experimental results we tested all of these metrics by applying the KNN strategy using them. We compare obtained results with the results of other classification methods presented in [6].

2. Nearest-neighbor classifier and distance metrics

In this section we give a short description of the nearest neighbor classifier and the distance metrics used in this paper.

2.1. Nearest neighbor classifier

Nearest neighbor classifier is a relatively simple and common used classification method which can be applied both for classification and regression. Since this method delays the process of modeling the training data until it is needed to classify test examples, this classifier is known as lazy learner. As a consequence, the efficiency of the KNN depends on the dimension of the training set (N_{tr})

and the number of attributes (N). For example, for each training vector the time complexity of 1-NN is $O(N_{tr}N)$.

The main principles of this method are based on finding all the training examples that are relatively similar to the attributes of the test example. Each example is represented as data point in n-dimensional space, where n is number of attributes. In the algorithm, distance (or similarity) between each test example $z = (x', y')$ and all the training examples $(x, y) \in D$ are calculated, in order to determine the nearest-neighbor list D_z. The k nearest neighbors of a given example z refer to the k training examples that are closest to z. These examples are further used to determine the class label of the test example. More precisely, once the nearest-neighbor list is obtained, the test example is classified based on majority class of its neighbors:

$$y' = \arg\max_{v} \sum_{(x_i, y_i) \in D_z} I(v = y_i),$$

where v is class label, y_i is class label for one of the nearest neighbors, and $I(\cdot)$ is an indicator function that returns value 1 if its argument is true and 0 otherwise.

The choice of the number k can significantly influence on the success of the classification. In some cases, if k is too small, the nearest-neighbor classifier may be susceptible to overfitting because of noise in the training data. On the other side, if k is too large, the nearest-neighbor classifier may misclassify the test record because its list of nearest neighbors may include data points that are located far away from its neighborhood [7].

If KNN is used for solving binary classification tasks, odd values of k are usually used to avoid ties, i.e., two classes labels achieving the same score. In the KNN presented in this paper, k takes each value from the set $\{1, 3, 5, 7, 9, 11, 13, 15\}$.

2.2. Distance metrics

As it is already mentioned, in cases when many nominal attributes are included in classification, standard metrics often can not be directly applied, since nominal attributes must be handled in a problem-specific way. In literature many distance functions for handling the nominal attributes are proposed. A detailed analysis of them is out of the scope of this paper and can be found for example in [8] and the references therein.

In this paper, for improving the classification process based on the KNN strategy, we decided to implement and test the following distance functions:

- Hamming-Euclidean overlap metric (HEOM)
- Frequency weighted overlap metric (FWOM)
- Heterogenous Valued difference metric (HVDM), actually three variants of this metric, slightly differing in the way of calculating the valued distance.

In addition, we implemented the numeric metric, which handles only the numeric attributes. We use this metric for calculating the distance between numerical attributes in HEOM, FWOM and HVDM metrics.

In the following subsections, we shortly describe the introduced metrics.

Numeric metric. Numeric metric ignores nominal attributes and bases classification model only on numerical ones. If $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two examples then

$$numeric(x,y) = \sqrt{\sum_{i=1}^{n} d_i^2(x_i, y_i)}.$$

In the considered dataset there are many NULL values assigned to numerical attributes. If in a pair of attributes only one NULL value appear, than we handle the problem in a simple, but effective way: the distance between the attribute x_i having a numeric value and the attribute y_i having NULL value is calculated as difference between x_i and the average value of the attribute i. In a case when both attributes are missing, than the distance is equal to 0. More precisely, the distance between two attributes is calculated by the following formula:

$$d_i(x_i, y_i) = \begin{cases} |x_i - y_i|, & x_i, y_i \neq NULL \\ |x_i - avg(i)|, & y_i = NULL \ \wedge \ x_i \neq NULL \\ |y_i - avg(i)|, & x_i = NULL \ \wedge \ y_i \neq NULL \\ 0, & \text{otherwise.} \end{cases}$$

$avg(i)$ is average value of i-th attribute.

Since the value $d_i(x_i, y_i)$ can be very large, it is divided by 4 standard deviations to scale value into a range that is usually of width 1. The numeric features are therefore normalized with $d(x_i, y_i) = \dfrac{|x_i - y_i|}{4\sigma}$, where σ is standard deviation.

HEOM metric. Hamming-Euclidean metric is a heterogeneous metric that use different attributes distance function on different kinds of attributes. This metric is introduced by Wilson and Martinez [8] and it is the combination of the Euclidean and Hamming metric. For nominal attributes, the Hamming distance is considered: the distance is equal to 0 if two attributes are equal and 1 if they are different or one of them is NULL. If attributes are numerical, the HEOM metric uses the Euclidean distance, which is similar as the distance calculated in the numeric metric.

Formally,

$$d_i(x_i, y_i) = \begin{cases} Hamming\ distance & \text{, if } i\text{-th attribute nominal;} \\ Euclidean\ distance & \text{, if } i\text{-th attribute numeric.} \end{cases}$$

Euclidean distance for attributes x_i and y_i is calculated as $|x_i - y_i|$. Similarly to the case of numeric metric, $d_i(x_i, y_i)$ can be very large, so numeric features are normalized by the formula $d(x_i, y_i) = \dfrac{|x_i - y_i|}{4\sigma}$, where σ is standard deviation.

Finally, if $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two examples then

$$heom(x,y) = \sqrt{\sum_{i=1}^{n} d_i^2(x_i, y_i)},$$

where $d_i(x_i, y_i)$ is Hamming-Euclidean distance.

FWOM metric. In HEOM metric all attributes have identical contributions to the overall distance. One way to control the influence of attributes is introducing different weights on different attributes. This approach is applied in the frequency weighted overlap metric (FWOM). The FWOM metrics is introduced in [9] and has a similar definition as HEOM metric, but the nominal attributes are assigned the appropriate weights, defined as

$$\omega_i = \frac{F(x_i) + F(y_i)}{F(x_i)F(y_i)}$$

where $F(x_i)$ and $F(y_i)$ are the frequencies of the attributes x_i and y_i in training data.

So, if $x = (x_1, x_2, ..., x_n)$ and $y = (y_1, y_2, ..., y_n)$ are two examples then

$$fwom(x,y) = \sum_{i=1}^{n} d(x_i, y_i),$$

where

$$d_i(x_i, y_i) = \begin{cases} \omega_i \cdot Hamming\ distance & \text{, if } i\text{-th attribute nominal;} \\ Euclidean\ distance & \text{, if } i\text{-th attribute numeric.} \end{cases}$$

HVDM metric. In HVDM metric, the valued difference metric instead Hamming metric is used for determine distance between nominal values. Valued difference metric is is defined as [8]:

$$HVDM(x,y) = \sqrt{\sum_{i=1}^{n} d_i^2(x_i, y_i)}$$

where n is number of attributes.

The function $d_i(x,y)$ returns a distnace between the two values x and y for attribute i and is defined as:

$$d_i(x,y) = \begin{cases} 1 & \text{, if } x \text{ or } y \text{ unknown;} \\ normalized\_vdm_i(x,y) & \text{, if } i \text{ is nominal;} \\ normalized\_diff_i(x,y) & \text{, if } i \text{ is numerical.} \end{cases}$$

The function $normalized\_diff_i$ is defined similarly to the previous cases when numerical attributes figure:

$$normalized\_diff_i(x,y) = \frac{|x-y|}{4\sigma_i}.$$

In order to deeper analyze the behaviour of the HVDM metric applied to the considered dataset, we considered three variants of calculating the distance between two records:

$$N1 : normalized\_vdm1_i(x,y) = \sum_{c=1}^{C} \left| \frac{N_{i,x,c}}{N_{i,x}} - \frac{N_{i,y,c}}{N_{i,y}} \right|,$$

$$N2 : normalized\_vdm2_i(x,y) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{i,x,c}}{N_{i,x}} - \frac{N_{i,y,c}}{N_{i,y}} \right|^2},$$

$$N3 : normalized\_vdm3_i(x,y) = \sqrt{C * \sum_{c=1}^{C} \left| \frac{N_{i,x,c}}{N_{i,x}} - \frac{N_{i,y,c}}{N_{i,y}} \right|^2},$$

where $N_{i,x}$ is the number of records in the training set that have value x for attribute i, $N_{i,x,c}$ is the number of records in the training set that have value x for attribute i and output class c. C is the number of output classes in the problem domain.

The difference between $N1$ and $N2$ is similar to a difference between Manhattan and Euclidean distance, while $N3$ is the function used in [10], where HVDM was first introduced. Using VDM, the average value for $N_{a,x,c}/N_{a,x}$ (as well as for $N_{a,y,c}/N_{a,y}$) is $1/C$. Since the difference is squared and then added C times, the sum is usually in the neighborhood of $C(1/C^2) = 1/C$. This sum is therefore multiplied by C to get it in the range $0,...,1$, making it roughly equal in influence to normalized numeric values.

3. Experimental results

This section contains experimental results obtained by application of the KNN classification algorithm to the chosen dataset of prokaryotic organisms.

All the tests are executed on the Intel i3-4000M CPU @2.4GHz with 12GB RAM under 64-bit Windows 10 Operating system. For each execution, only one thread/processor is used. The KNN algorithm is implemented in C programming language and compiled with Visual Studio 2012 compiler.

For each problem dataset, KNN is executed multiple times by selecting K from the set $\{1,3,5,7,9,11,13,15\}$ and by selecting different distance metric from the set {HEOM, FWOM, HVDM1, HVDM2, HVDM3, Numeric}.

Initially, each problem dataset is randomly separated to two parts: the first one is called the training subset and it consists of about 70% of records from the whole problem dataset, while the remaining 30% of records (test subset) is used to test the quality of KNN for a selected K and distance metric. Test accuracy is calculated as a percentage of accurately assigned classes to feature vectors from the test set. During this class assignment, only nearest neighbours from the training subset are considered.

3.1. Dataset collection

The data used in this work refer to the prokaryotic organisms. The data are extracted from the NCBI (National Center for Biotechnology Information) site (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, as of February 9th, 2012). Later, some characteristics of organisms were added from the Patric (http://patricbrc.org) and Doe databases (http://img.jgi.doe.gov/). All data were stored in the table entitled Characteristics of Organisms. That table contains 1971 different records and for each record there are total of 30 attributes. In this research

TABLE 1: Considered classifications

Class.	Attributes	Target class	#nominal att.
Class 1	shape, organism size, arrangement	superkingdom	2
Class 2	shape, organism size, arrangement	phylum	2
Class 3	shape, motility, endospores	superkingdom	3
Class 4	shape, motility, endospores	phylum	3
Class 5	habitat, temp range, optimal temp	superkingdom	2
Class 6	habitat, temp range, optimal temp	phylum	2
Class 7	temp range, optimal temp	habitat	1
Class 8	pathogenic, oxygenreq, optimal temp	superkingdom	2
Class 9	pathogenic, oxygenreq, optimal temp	phylum	2
Class 10	oxygenreq, optimal temp	pathogenic	1
Class 11	habitat, motility	superkingdom	2
Class 12	habitat, motility	phylum	2

10 attributes are used: shape, organism size, motility, habitat, optimal temp, arrangement, endospores, pathogenic, oxygenreq and temperature range.

In the Table 1 we show an overview of 12 classifications analyzed in this work. The first column contains labels of classifications, the second column is the list of attributes which are used in particular classification, the third is the target class and in the forth column we show the number of nominal attributes used in the classification. For example, the first classification uses attributes: shape, organism size and arrangement. The target class is superkingdom (which can take values Bacteria or Archea). Attributes shape and arrangement are nominal, organism size is numerical, so there are two nominal and one numerical attribute in this classification. The third classification uses three nominal attributes (shape, motility, endospores) to form classification model for same target class as the first one.

3.2. Results of the classifications

In the Tables 2-13 the results obtained on these 12 classifications are shown. The first column of each table contains the number of considered neighbours, the second column contains the result obtained by HEOM metric and the third one by FWOM metric. The next three columns contain results obtained by the HVDM metrics (HVDM1, HVDM2, HVDM3 respectively), and the last column contains results obtained by using the numeric metric. The best obtained result is bolded. Since all attributes in classifications 3, 4, 11 and 12 are nominal, numeric metric can not be applied to them.

From the Table 2 (classification 1) one can see that the best score of correctly classified test data is obtained by the HEOM for 11 neighbors and with two variants of the HVDM metrics (HVDM1 and HVDM2) for 13 neighbors. The weakest results are obtained by the numeric metric. From the Table 3 it can be seen that the best result is obtained by the HEOM metric and 11 neighbors. Numeric metric again gives the weakest results, which indicates that the proposed distance metrics improves the success of the classification. In the third classification (Ta-

TABLE 2: Class. 1: Target class Superkingdom

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	90.88%	91.39%	90.71%	90.71%	90.71%	89.70%
3NN	94.76%	94.93%	94.93%	94.93%	94.76%	92.57%
5NN	95.44%	95.44%	95.78%	95.78%	95.78%	94.09%
7NN	95.27%	95.10%	95.27%	95.27%	95.27%	93.92%
9NN	95.78%	95.61%	95.61%	95.61%	95.61%	94.26%
11NN	**95.95%**	95.78%	95.78%	95.78%	95.61%	94.43%
13NN	95.44%	95.27%	**95.95%**	**95.95%**	95.27%	94.26%
15NN	95.44%	95.27%	95.44%	95.44%	95.44%	94.26%

TABLE 3: Class. 2: Target class Phylum

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	52.20%	53.21%	40.20%	40.20%	40.20%	41.39%
3NN	56.42%	56.42%	43.92%	43.92%	43.92%	43.41%
5NN	59.29%	59.12%	50.00%	50.00%	50.00%	48.48%
7NN	58.95%	58.61%	50.51%	50.51%	50.51%	49.16%
9NN	60.47%	60.14%	51.69%	51.69%	51.69%	50.84%
11NN	**60.64%**	60.14%	51.86%	51.86%	51.86%	51.01%
13NN	60.47%	60.30%	52.53%	52.53%	52.53%	50.34%
15NN	58.45%	58.45%	51.18%	51.18%	51.18%	51.35%

TABLE 4: Class. 3: Target class Superkingdom

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	93.41%	93.41%	93.41%	93.41%	93.41%	-
3NN	76.01%	76.01%	76.01%	76.01%	76.01%	-
5NN	**94.93%**	**94.93%**	**94.93%**	**94.93%**	**94.93%**	-
7NN	**94.93%**	**94.93%**	**94.93%**	**94.93%**	**94.93%**	-
9NN	**94.93%**	**94.93%**	**94.93%**	**94.93%**	**94.93%**	-
11NN	**94.93%**	**94.93%**	**94.93%**	**94.93%**	**94.93%**	-
13NN	**94.93%**	**94.93%**	**94.93%**	**94.93%**	**94.93%**	-
15NN	**94.93%**	**94.93%**	**94.93%**	**94.93%**	**94.93%**	-

TABLE 5: Class. 4: Target class Phylum

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	54.39%	54.39%	42.91%	42.91%	42.91%	-
3NN	52.03%	52.03%	42.91%	42.91%	42.91%	-
5NN	51.86%	51.86%	31.93%	31.93%	31.93%	-
7NN	55.91%	55.91%	31.93%	31.93%	31.93%	-
9NN	55.74%	55.74%	42.91%	42.91%	42.91%	-
11NN	56.93%	56.93%	42.91%	42.91%	42.91%	-
13NN	**58.78%**	**58.78%**	42.91%	42.91%	42.91%	-
15NN	**58.78%**	**58.78%**	42.91%	42.91%	42.91%	-

TABLE 6: Class. 5: Target class Superkingdom

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	95.61%	95.27%	95.27%	95.27%	95.61%	94.59%
3NN	96.11%	96.11%	96.11%	96.11%	96.11%	94.59%
5NN	96.11%	95.95%	95.95%	95.95%	95.95%	94.59%
7NN	96.79%	**97.13%**	96.62%	96.62%	96.62%	94.59%
9NN	96.79%	96.62%	96.62%	96.62%	96.62%	94.59%
11NN	96.79%	96.62%	96.62%	96.79%	96.62%	94.59%
13NN	96.79%	96.62%	96.62%	96.62%	96.62%	94.59%
15NN	96.45%	96.28%	96.62%	96.62%	96.62%	94.59%

TABLE 7: Class. 6: Target class Phylum

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	31.93%	31.93%	26.86%	26.86%	26.86%	39.19%
3NN	32.09%	32.09%	37.33%	37.33%	37.33%	44.59%
5NN	48.31%	**48.48%**	46.11%	48.14%	46.11%	44.59%
7NN	47.97%	47.97%	47.13%	47.13%	47.13%	44.59%
9NN	37.84%	38.01%	46.62%	46.62%	46.62%	44.59%
11NN	38.01%	38.68%	46.62%	46.62%	46.62%	44.59%
13NN	36.49%	37.33%	46.28%	46.28%	46.28%	44.59%
15NN	36.49%	36.99%	46.28%	46.28%	46.28%	44.59%

TABLE 8: Class. 7: Target class Habitat

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	29.73%	29.73%	29.73%	29.73%	29.73%	26.69%
3NN	30.57%	30.57%	30.57%	30.57%	30.57%	28.04%
5NN	**45.61%**	**45.61%**	45.44%	45.44%	**45.61%**	43.24%
7NN	44.76%	44.76%	44.76%	44.76%	44.76%	42.57%
9NN	44.76%	44.76%	44.76%	44.76%	44.76%	42.57%
11NN	44.93%	44.76%	44.93%	44.93%	44.93%	42.74%
13NN	31.93%	31.76%	31.93%	31.93%	31.93%	43.24%
15NN	30.91%	30.74%	30.91%	30.91%	30.91%	28.55%

TABLE 9: Class. 8: Target class Superkingdom

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	94.59%	94.59%	94.59%	94.59%	94.59%	94.59%
3NN	94.59%	94.59%	94.59%	94.59%	94.59%	94.59%
5NN	94.59%	94.59%	94.59%	94.59%	94.59%	95.78%
7NN	94.59%	94.59%	94.59%	94.59%	94.59%	94.59%
9NN	94.59%	**95.95%**	94.59%	94.59%	94.59%	94.59%
11NN	94.59%	94.59%	94.59%	94.59%	94.59%	94.59%
13NN	94.59%	94.59%	94.59%	94.59%	94.59%	95.78%
15NN	94.59%	95.10%	94.59%	94.59%	94.59%	94.59%

ble 4) all methods obtain similar results. In the fourth classification the best results are obtained by HEOM and FWOM metrics, for larger k, ($k = 13$ and $k = 15$). For the classification 5, all methods obtains similar results and the best one if reached by the FWOM metric and $k = 7$. In the classification 6, HVDM metrics are more successful in average, but the best result is again obtained by the FWOM metric and $k = 5$. In the classification 7 best results are obtained for $k = 5$ (HEOM, FWOM and HVDM3). In the classification 8 all methods achieves good and similar results and the best one is obtained by the FWOM metric and $k = 9$. In classification 9, FWOM and HEOM metrics obtains best results for larger values of k. In classification 10 and 11 all methods obtain similar results. In general, better results are obtained by larger values of k. In the last classification, best results are obtained by HEOM and FWOM metrics and $k = 15$.

3.3. The comparison with previous methods

Table 14 contains the results obtained by several classification algorithms: Sprinter from IBM Intelligent Miner package which is based on Decision Tree algorithm, CHAID from IBM SPSS Statistics 23 (SPSS) which is also based on the Decision tree algorithm, Nave Bayes algorithm from WEKA package and Jrip algorithm also from WEKA which is Rule-Based Classifier. All these results are extracted from [6]. The best result of each classification is bolded.

From the Table 14 it is evident that presented metrics can improve the NN strategy for classifications in $7/12$ cases. The presented strategy improves results for classifications 1-5, 8 and 11, achieving better results than those presented in [6]. The proposed methods can be applied to classification containing both numerical and nominal attributes. From a deeper analysis of the obtained data, one can conclude that HEOM and FWOM metrics behave similary. As expected, since HVDM 1-3 metrics are defined in a similar way, the obtained results are also similar.

4. Conclusions and future work

In this paper we presented several distance metrics that can be used for classifications which involve nominal attributes. In cases when many nominal attributes appear, standard classification tools usually ignore their appearance, causing inaccurate and unreliable results. In order to overcome this problem, we introduced several distance metrics that can be applied to the considered classifications.

Experimental results indicate a high reliability of the proposed methods. This strategy improves previously known results in seven of total of twelve cases. The obtained results indicate that this approach can be used for classification of such datasets.

This research can be extended in several ways. For example, the proposed algorithms can be applied to other biological datasets. In classifications where more attributes are considered, the proposed KNN approach can be combined

TABLE 10: Class. 9: Target class Phylum

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	27.36%	27.36%	28.21%	28.21%	28.21%	39.19%
3NN	27.20%	27.20%	29.39%	29.39%	29.39%	44.59%
5NN	33.45%	33.45%	36.82%	36.82%	36.82%	44.59%
7NN	44.09%	44.09%	44.59%	44.59%	44.59%	44.59%
9NN	44.09%	44.09%	44.59%	44.59%	44.59%	44.59%
11NN	43.92%	43.92%	37.84%	37.84%	37.84%	44.59%
13NN	**46.28%**	**46.28%**	44.59%	44.59%	44.59%	44.59%
15NN	**46.28%**	**46.28%**	44.59%	44.59%	44.59%	44.59%

TABLE 11: Class. 10: Target class Pathogenic

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	46.45%	46.45%	46.28%	46.28%	46.28%	47.97%
3NN	73.82%	73.82%	73.31%	73.31%	73.65%	67.06%
5NN	75.84%	75.84%	75.68%	75.68%	75.84%	66.22%
7NN	44.93%	44.93%	44.93%	44.93%	44.43%	48.65%
9NN	**77.20%**	**77.20%**	**77.20%**	**77.20%**	77.03%	66.55%
11NN	**77.20%**	**77.20%**	**77.20%**	**77.20%**	**77.20%**	66.55%
13NN	**77.20%**	**77.20%**	**77.20%**	**77.20%**	**77.20%**	66.55%
15NN	**77.20%**	**77.20%**	**77.20%**	**77.20%**	**77.20%**	66.55%

TABLE 12: Class. 11: Target class Superkingdom

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	84.29%	84.29%	84.29%	84.29%	84.29%	-
3NN	92.57%	92.57%	92.57%	92.57%	92.57%	-
5NN	92.57%	92.57%	92.57%	92.57%	92.57%	-
7NN	93.75%	93.75%	93.75%	93.75%	93.75%	-
9NN	**94.43%**	**94.43%**	**94.43%**	**94.43%**	**94.43%**	-
11NN	**94.43%**	**94.43%**	**94.43%**	**94.43%**	**94.43%**	-
13NN	**94.43%**	**94.43%**	**94.43%**	**94.43%**	**94.43%**	-
15NN	**94.43%**	**94.43%**	**94.43%**	**94.43%**	**94.43%**	-

TABLE 13: Class. 12: Target class Phylum

	HEOM	FWOM	HVDM1	HVDM2	HVDM3	Numeric
1NN	43.41%	43.41%	42.91%	42.91%	42.91%	-
3NN	44.09%	44.09%	10.64%	42.91%	42.91%	-
5NN	41.22%	41.22%	10.64%	10.64%	10.64%	-
7NN	41.72%	41.72%	42.91%	10.64%	10.64%	-
9NN	44.26%	44.26%	42.91%	42.91%	42.91%	-
11NN	45.95%	45.95%	42.91%	42.91%	42.91%	-
13NN	45.95%	45.95%	42.91%	42.91%	42.91%	-
15NN	**48.14%**	**48.14%**	42.91%	42.91%	42.91%	-

TABLE 14: Comparative results obtained by different classification algorithms

Class	Sprinter	CHAID	Naïve Bayes	Jrip	KNN	Improved NN
class1	93.75%	92.70%	92.56%	93.74%	90.37%	**95.95%**
class2	46.40%	51.10%	-	53.98%	49.16%	**60.64%**
class3	21.07%	93.30%	92.72%	94.25%	-	**94.93%**
class4	1.00%	56.50%	53.13%	54.48%	-	**58.78%**
class5	83.79%	96.60%	94.25%	95.77%	92.51%	**97.13%**
class6	4.00%	47.50%	-	43.82%	**52.02%**	48.48%
class7	8.00%	52.20%	49.07%	47.20%	**55.35%**	45.61%
class8	87.23%	94.00%	93.06%	94.59%	91.57%	**95.95%**
class9	9.00%	44.30%	0.00%	43.32%	**49.09%**	46.28%
class10	48.34%	**83.60%**	64.47%	82.24%	76.65%	77.20%
class11	55.83%	92.70%	92.22%	92.22%	-	**94.43%**
class12	0.00%	47.30%	**48.90%**	46.02%	-	48.14%

with a feature selection algorithm. It would be also interesting to analyse other distance metrics that can be adopt for classifications of prokaryotic organisms.

References

1. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten: The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1, (2009)
2. M.R. Berthold, N. Cebron, F.Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, B. Wiswedel: KNIME: The Konstanz Information Miner, Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer, (2007)
3. IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.
4. IBM DB2 Intelligent Miner for Data, Using the Intelligent Miner for Data, First Edition, (2002)
5. Grbić M. "Analysis of classification algorithms applied to some prokaryotic organisms" (poster), The Ninth International Biocuration conference, Geneve, 2016.
6. Grbić M. "Grouping organisms by various classification methods depending on genotype and phenotype characteristics", Master thesis (in Serbian), Faculty of Mathematics, Belgrade, 2016.
7. Tan, Pang-Ning and Steinbach, Michael and Kumar, Vipin: Introduction to Data Mining. Pearson. (2006)
8. Wilson D., Martinez T.:: Improved Heterogeneous Distance Functions. Journal of Aritificial Intelligence Research. (1997)
9. Huang: A fast clustering algorithm to cluster very. In Reaserch Issues on Data Mining and Knowledge Discovery. (1997)
10. Wilson D., Martinez T.: Heterogeneous Radial Basis Function Networks Proceedings of the International Conference on Neural Networks (1996)

T-cell epitope prediction, the influence of amino acids physicochemical propeterties and frequencies on identifying MHC binding ligands

Davorka R. Jandrlić[1], Nenad S. Mitić[2], and Mirjana D. Pavlović[3]

[1] Faculty of Mechanical Engineering, Kraljice Marije 16, Belgrade, Serbia
djandrlic@mas.bg.ac.rs
[2] Faculty of Mathematics, University of Belgrade, Studentski trg 16
11000 Belgrade, Serbia
nenad@matf.bg.ac.rs
[3] Institute of General and Physical Chemistry, Studentski trg 12/V, Belgrade, Serbia
11000 Belgrade, Serbia
mpavlovic@iofh.bg.ac.rs

Abstract. Binding of peptides to MHC class I molecules is essential and the most selective step that determines T cell epitopes. Therefore, prediction of MHC-peptide binding presents the principal basis for anticipating potential T cell epitopes. The immense relevance of epitope identification in vaccine design has prompted the development of many computational methods. All of them have advantages and drawbacks. Although some available methods have reasonable accuracy, there is no guarantee that all models produce good quality predictions [1]. The aim of computational methods is to reduce the laboratory expensive experiments [2], that is way every effort to improve performance of existing methods or make reliable new method is important.

Keywords: bioinformatics, data mining, MHC binding prediction, *k*-mean clustering, SVM

1. Introduction

Bioinformatics approaches play a critical role on analyzing multiple genomes to select the protective epitopes in silico. It is conceived that cocktails of defined epitopes or chimeric protein arrangements, including the target epitopes, may provide a rationale design capable to elicit convenient humoral or cellular immune responses [4]. Bioinformatics tools and immunological software are necessary in order to facilitate the design and development of vaccines. Predictions obtained by the use of different parameters and methods can greatly improve the accuracy compared to accuracy of a single method. During the development of such tools [5] that support bioinformatic research related to prediction of T cell epitopes, protein hydropathy, disordered and disordered binding regions, and applying them to a large number of proteins [8, 9], there was evidence that there were some characteristics that influence the binders appearance. We examined a new approach for identifying the most relevant physicochemical properties (PC), for classification of petides into MHC-binding ligands or non

binding ligands [7]. The new methods were developed that take into account the physicochemical properties of amino acids and their frequencies. The developed classification models are rule based and use k-means clustering technique for extracting the most important properties. The obtained results indicate that the physicochemical properties of amino acids contribute significantly to the peptide-binding affinity and that the different alleles are characterized by a different set of the physicochemical properties. Results from these models are used as input features to two machine learning models, based on support vector machine technique for classification and regression problem [6]. The models for quantitatively and qualitatively predicting MHC-binding ligands, were made. The resulting models have shown comparable performance, or in some cases better than two of the currently best available predictors: *NetMHCpan* and SMM^{PMBEC} [3]. The new models could be used as complement to the best existing methods.

2. Materials and methods

The Immune Epitope Database (IEDB) (http://www.iedb.org/), June 2015 version, served as data source. The research has been limited to peptides of 9 amino acids (AAs) in length because nonamers are the most common MHC-I epitopes, and to peptides of 15 amino acids (AAs) in length for MHC-II class, because there are only enough experimental data to construct good models for that length of the peptide (see section Materials and methods in [7, 6] for details).

2.1. The rule based classification models

The development of these models is aimed at identifying the main features of the peptides that separate binders from non binders. In this phase, the peptide is represented using its combination of unigrams and bigrams frequencies and specific PC properties, as described below. Calculating the frequency of AAs at appropriate positions in a peptide is aimed at extracting the features of occurrence of AAs in peptide binders and non-binders, which would enable easier classification. Instead of the standard calculation of AA frequency by position in a peptide, we used a modified calculation of frequency which was successfully implemented in document classification [10].

Equation	Definition
$df(t,S)$	Frequency of the term t in a set of peptides S.
$idf(t,S) = log_2(\|S\|)/(df(t,S))$	Inverse frequency of the term t in the set of peptides S.
$tf(t,Peptide)$	The number of occurrences of the term t in the *Peptide*

TABLE 1: The frequency measures

In the example of peptide $p = LVIKALLEV$, t could be from the set {L, V, I, K, A, L, L, E, V, L, V, VI, IK, KA, AL, LL, LE, EV }. The issues of non-linearity of

AA frequency and peptides that have AAs or bigrams that do not occur in both classes are resolved with the introduction of smoothing factors [11]:

$$\Delta tfidf(t_i, Peptide, S^+, S^-) = tf(t_i, Peptide) * log_2 \frac{|S^+|}{df(t_i, S^+)} * \frac{df(t_i, S^-)}{|S^-|} \quad (1)$$

$$\Delta BM25idf(t_i, Peptide, S^+, S^-) = log \frac{(|S^+| - \Delta tfidf(t_i, S^+) + 0.5) * tfidf(t_i, S^-) + 0.5}{(|S^-| - \Delta tfidf(t_i, S^-) + 0.5) * tfidf(t_i, S^+) + 0.5} \quad (2)$$

- t_i is AA or bigram in peptide *Peptide* from set S, at position i
- S^+ and S^- are the subsets of S, of positive ligands (binders) and negative ligands (non binders), respectively
- $|S^+|$ and $|S^-|$ are the cardinalities of the positive and negative sets.

If peptides are 9 AAs in length, an assigned vector is 9 + 8 = 17. The 119 PC properties were taken from [12], with the aim to investigate specific allele characteristics and to evaluate influence of each individual PC property on classification peptides into MHC binders or non-binders. The peptide is firstly encoded with single PC property, in this way peptide is represented with vector of length 17 (9 + 8) with the numerical value obtained by applying PC property on appropriate consecutive AAs and bigrams from that peptide. This procedure is carried out for every single PC. To evaluate the importance of single PC property, for each PC property fk (k = 1, ...,119) and every single allele, a new rule based classification model was constructed. This procedure is explained in details in [7]. The scheme of peptide encoding that combines the frequencies an the pc characteristics is illustrated in Fig. 1

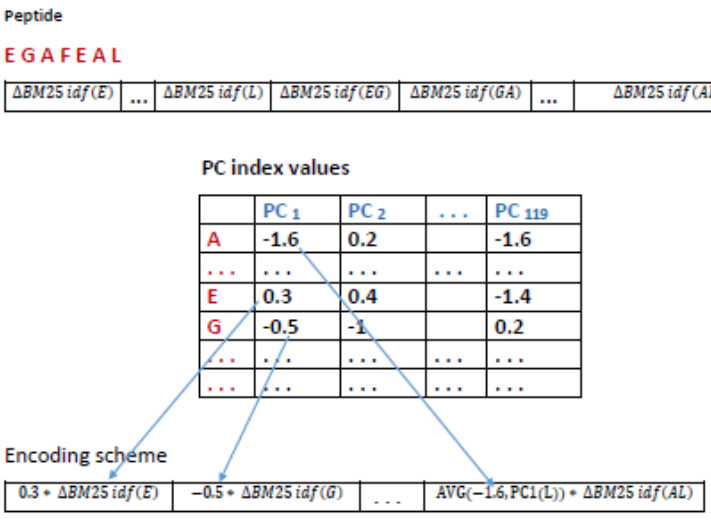


FIG. 1: Encoding scheme

The development of these rule based models was intended to determine the importance of each of the considered properties in the separation binders from

non binders. The most important properties are identified and incorporated into new machine learning models described bellow.

2.2. The machine learning models

Some of the existing methods and models, taken from recently review paper [13] are listed in the table 2. In existing methods the most conventional are

TABLE 2: Existing machine learning methods

Name	Method	Scheme
ANNPred[4]	ANN	Sparse encoding
nHLAPred[5]	ANN/PSSM	Sparse encoding
Zhu et al.,	Decision tree	N/A
KISS[6]	SVM	Heckerman et al.
POPI[7]	SVM	Physicochemical properties
SVMHC[8]	SVM	Sparse encoding
NetMHC[9]	ANN	Sparse encoding/BLOSUM50
NetMHCpan[10]	ANN	Sparse encoding/BLOSUM50
MHCPRED[11]	QSAR regression	-
SVRMHC[12]	SVM	Sparse encoding/11 pc

sparse encoding strategies, where each AA in a peptide is encoded as a 20-digit binary number, a single 1 and 19 zeros. The second is BLOSUM50 encoding, whereby AAs are encoded as the BLOSUM50 score for replacing the AA with each of the 20 AAs , or a combination of these two strategies.
Here, the peptide is represented using three new types of features, obtained by the extensive training and testing of the rule based models and the application of the knowledge gained in the previous step. The combination of these features served as input for three different support vector machine (SVM) and support vector regression (SVR) models. SVM models were created for binary classification, i.e. to predict whether a peptide is or is not an epitope, while the SVR models were made for predicting the binding affinity of a peptide to a particular allele. The weighting schemes are based on:

- position-dependent amino acid frequencies (Δ-BM25-IDF),
- BLOSUM and VOGG substitution of amino acids,
- physicochemical properties of amino acids (the 10 best PC) and
- molecular properties of amino acids ($z5$ descriptors)

The first encoding scheme. The first encoding scheme is combination of BLOSUM (VOGG) and Δ-BM25-IDF encoding (Fig. 2).

The second encoding schema. This encoding strategie is constucted using 10 best PC properties for unigrams and bigrams within peptide.

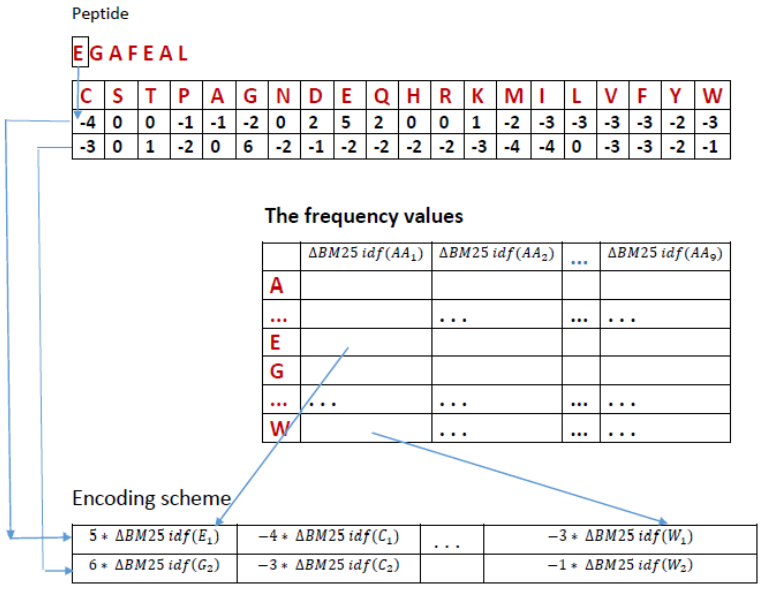


FIG. 2: Encoding scheme based on BLOSUM and Δ-BM25-IDF

The third encoding scheme. This scheme is expansion of the first scheme with molecular descriptors. We combined three different encoding schemes of BLOSUM62 (VOGG) encoding, Δ-BM25-IDF weighting and Z5-descriptors. As each AA in a peptide is replaced by the appropriate type from the VOGG matrix, the peptide is represented by a vector 180 in length. Every component of the resulting vector is multiplied by the Δ-BM25-IDF weight obtained for the AA it represents. Every AA in a peptide is represented with another 5 descriptors, which means that another 9 x 5 = 45 components are added to the vector. Finally, the peptide is represented by a vector 225 in length, whereby we have covered the frequency of AAs, the composition of the PC properties that best describe the molecular characteristics of the AAs and possible substitution.

All three schemes are described in detail in the paper [6], as well as the results obtained by the use of these models.

Models building. Support vector machine (SVM) is assumed to be a very powerful algorithm that often achieves superior classification performance in comparison with other classification algorithms. They are efficient enough to handle very large-scale classification in both number of samples and number of features which was the case here. SVM method was used for two class problem (binary classification of peptides into epitopes and non epitopes) and to predict binding affinity of the peptides. For that purpose an affinity is logarithmically scaled to continuous numerical value from the interval [0, 1]. The basic idea of SVMs is to

map data of samples into a high dimensional Hilbert space and to seek a separating hyperplane in this space (separates the positive from negative examples). The kernel functions that where used to perform the non-linear mapping into feature space are:

- Radial Basis Function (RBF) for regression problem:
  $K(x_i, x_j) = e^{-\gamma|x_i - y_j|^2}$
- Polynomial Function for binary classification:
  $K(x_i, x_j) = (x_i x_j + 1)^e, e = 1, 2,$

To evaluate the performance of our methods comprehensively, we report standard performance measures, including **accuracy**, **precision**, **recall**, **AROC**, **F-measure**, **Pearsons CC** and **Kappa statistic**. The experiments, for all models, were performed with a 10-fold cross-validation (CV) on the training set (70%). The remaining data (30%) were further used as a blind test for assessing models obtained in this way. The parameters were optimized for both techniques with a **grid search algorithm** $(C, \gamma, exponent)$.

3. Results

Prediction performance for all models are presented in table 3. Taking the average value of all measures as an assessment of the quality of all three models, it appears that the third model gives the best results (in the case of binary). Still, the results of the models based on these three schemes are quite close; all of them provide good results. It can be concluded that the calculation of inverse frequency of amino acids and encoding of the peptide by measures effectively used for text classification produce excellent results when used in combination with BLOSUM encoding for the classification of epitopes and prediction of the binding affinity for MHC-I proteins. The results of the second model indicate that the PC properties play an important role in the binding of peptides to MHC-I molecules. It was the results obtained with models based on the second scheme of encoding that motivated the creation of the third scheme, which included the encoding of peptides with Z-descriptors, because Z-descriptors actually represent a certain combination of PC and molecular properties (unique for all alleles, unlike the case in our second model).

TABLE 3: Performance of SVM and SVR models for all three schemas (AVG values)

SVM								SVR	
	Accuracy		Permormance on test set 10 cv					CC	
Allele	10-fcv	Test set	Precision	Recall	*F*-measure	*AROC*	Kappa	10-fcv	Test
AVG scheme 1	84.75	85.54	0.86	0.86	0.85	0.86	0.71	0.75	0.74
AVG scheme 2	87.93	87.47	0.89	0.89	0.89	0.87	0.83	0.80	0.80
AVG scheme 3	88.17	87.66	0.88	0.88	0.88	0.86	0.72	0.78	0.78

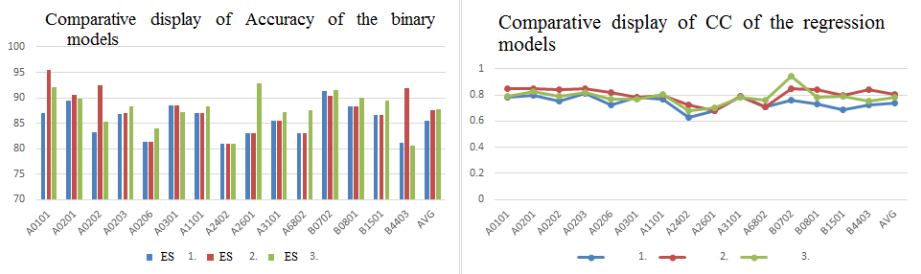


FIG. 3: SVM and SVR models comparative results

Performance comparison with other methods. Performance comparison of our three models with SMMPMBEC and NetMHCpan predictors

TABLE 4: The comparative results of the existing predictors SMMPMBEC, NetMHCpan and new developed regression models

Allel	SMMPMBEC	NetMHCPan	Schema 1	Schema 2	Schema 3
0101	0.66- 0.63	0.82 -0.77	0.80-0.78	**0.85-0.85**	0.75-0.79
A0201	0.80-0.78	0.869 -0.85	0.79-0.80	0.84-0.85	0.82-0.82
A0202	0.79 -0.80	0.83-0.84	0.79-0.75	0.81-0.84	0.78-0.79
A0203	0.83-0.79	0.86-0.88	0.79-0.81	0.84-0.85	0.81-0.82
A0206	0.75 -0.70	0.80- 0.74	0.67-0.72	**0.82-0.82**	0.76-0.77
0301	0.79- 0.79	0.81- 0.83	0.76-0.78	0.76-0.78	0.79-0.77
1101	0.78- 0.78	0.86-0.85	0.76-0.77	0.82-0.80	0.81-0.80
2402	0.67- 0.74	0.71-0.75	0.64-0.63	0.75-0.72	0.68-0.68
2601	0.62-0.59	0.78-0.76	0.77-0.68	0.77-0.68	0.70-0.70
3101	0.77- 0.72	0.83-0.73	0.77-0.79	0.77-0.79	0.81-0.78
6802	0.71-0.81	0.80-0.88	0.70-0.71	0.70-0.71	0.80-0.76
B0702	0.72- 0.78	0.84- 0.81	0.79-0.76	0.79-0.85	0.79-0.94
B0801	0.66- 0.81	0.76-0.89	0.74-0.73	**0.84-0.84**	**0.80-0.78**
B1501	0.68- 0.74	0.74- 0.809	0.70-0.69	**0.80-0.80**	0.77-0.79
B4403	0.75-0.81	0.80-0.83	0.71-0.72	**0.85-0.84**	0.77-0.75

4. Conclusion

The models presented here provide good results. The calculation of frequency of amino acids and using the measures effectively used for text classification produce excellent results for epitope classification. The results of models based on the PC properties indicate that the PC properties play an important role in the binding of peptides to MHC-I molecules and have direct influence on binding affinity. All the results point out the features of peptides that can be used to

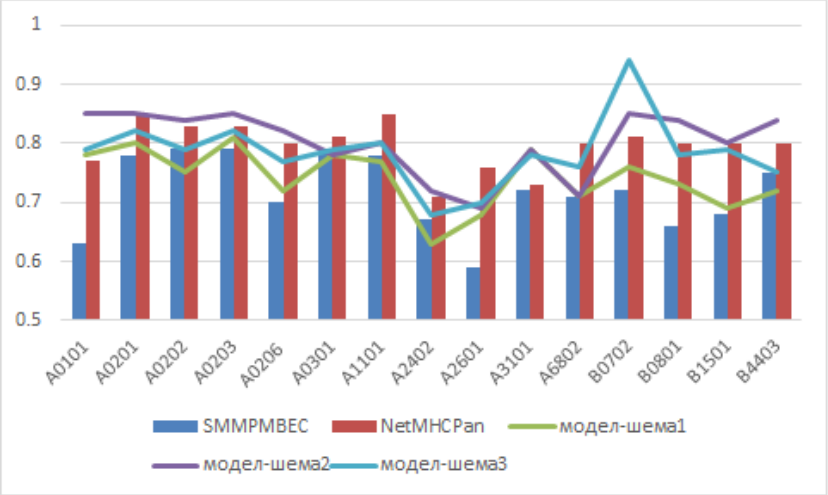


FIG. 4: Comparative results of SVR models with other methods

more easily identify potential epitopes, and suggest possible future direction for improving existing predictors.

Acknowledgments

This work was supported in part by Ministry of Education, Science and Technological Development of the Republic of Serbia, projects No. 174021, 174002, and III44006.

References

1. Brusić V, Bajić VB, Petrovsky N.,*Computational methods for prediction of T-cell epitopes–a framework for modelling, testing, and applications*, *Methods, 34(4), 436-443*, (2004)
2. Irini A. Doytchinova, Pingping Guan and Darren R. Flower, *EpiJen: a server for multistep T cell epitope prediction*, *BMC Bioinformatics*, (2006)
3. Nielsen M., Zhang H., Lundegaard C., *Pan-specific MHC class I predictors: a benchmark of HLA class I pan-specific prediction methods*, *Bioinformatics*, (2009)
4. https://doi.org/10.1016/j.jbi.2014.11.003, (2015)
5. D. R. Jandrlić and G. M. Lazi and N. S. Miti and M. D. Pavlovi, *Software tools for simultaneous data visualization and T cell epitopes and disorder prediction in proteins*, *Journal of Biomedical Informatics*, (2016)
6. D. R. Jandrlić, *SVM and SVR-based MHC-binding prediction using a mathematical representation of peptide sequences*, *Computational Biology and Chemistry*, (2016)
7. D. R. Jandrlić, *The rule based classsification models for MHC binding prediction and identification of the most relevant physicochemical properties for the individual allele*, *University thought - Publication in Natural Sciences*, (2016)
8. N. S. Mitić and M. D. Pavlović and D. R. Jandrlić, *Epitope distribution in ordered and disordered protein regions - Part A. T-cell epitope frequency, affinity and hydropathy*, *Journal of Immunological Methods*, (2014)

9. M. D. Pavlović and D. R. Jandrlić and N. S. Mitić , *Epitope distribution in ordered and disordered protein regions. Part B - Ordered regions and disordered binding sites are targets of T- and B- cell immunity*, *Journal of Immunological Methods*, (2014)
10. J. Martineau and T. Finin, *Delta TFIDF: An Improved Feature Space for Sentiment Analysis*, *In Proceedings of the Third AAAI International Conference on Weblogs and Social Media, San Jose, CA, May. AAAI Press.*, (2009)
11. T. Joachims, *A probabilistic analysis of the rocchio algorithm with tfidf for text categorization*, *InInternational Conference on Machine Learning (ICML)*, (1997)
12. F. Tian and L. Yang and F. Lv and et al., *In silico quantitative prediction of peptides binding affinity to human MHC molecule: an intuitive quantitative structure-activity relationship approach*, *Amino Acids*, (2009)
13. H Luo, H Ye, HW Ng, L Shi, W Tong, DL Mendrick et al., *Machine learning methods for predicting hlapeptide binding activity*, *Bioinformatics*, (2015)

Networks of Interaction in Moving Animal Groups and Collective Changes of Direction

Asja Jelić

The Abdus Salam International Centre for Theoretical Physics (ICTP),
Strada Costiera 11, 34014 Trieste, Italy
asja@ictp.it

Abstract. Animal groups on the move are a paradigmatic example of collective behaviour in social species. The most striking features of this collective motion are sudden coherent changes in the travel direction of the whole group. Such a coordinated collective behaviour requires fast and robust transfer of information among individuals in order to prevent cohesion loss. However, little is known about the mechanism by which natural groups achieve this robustness. Furthermore, collective directional switching often emerges not as a response to an external alarm cue, but spontaneously from the intrinsic fluctuations in individual behaviour. In particular, it is not yet clear the role of the underlying structure of the communication network in these events. In this paper, we present an overview of an experimental and theoretical study of spontaneous collective turns in natural flocks of starlings, which reveals the mechanisms behind this phenomena.

Keywords: collective animal behaviour, self-organization, decision-making

1. Introduction

Moving animal groups are a paradigm of collective behaviour in social species. Their collective motion is characterised by sudden, coherent changes in a travel direction of the whole group [1–9]. During this collective decision, all members of the group are required to go through a behavioural change of state. Little is known, however, about the mechanism that triggers these collective changes and enables fast and robust transfer of information across the whole system. Sometimes, collective directional changes occur as a response to an external alarm cue, such are evasive manoeuvres of animal groups under predatory attacks. Frequently, however, they arise spontaneously from the intrinsic fluctuations in individual behaviour [6–9]. In both of these cases, the collective change of state usually starts from a localized spatial origin – a few individuals that are close to each other. Once the decision to change direction is formed among the close-by group members, it propagates through the whole system and all individuals of the group change their direction. The ability of a group to perform such a coordinated behaviour without loss of cohesion crucially depends on fast and robust transfer of information among individuals.

In this paper, we review latest results of an experimental and theoretical study of spontaneous collective turns in natural flocks of starlings [8–13]. Employing a recently developed tracking algorithm, we were able to reconstruct

three-dimensional trajectories of each bird in a flock for the whole duration of a tracking event [10]. Having these detailed data enabled us to analyse the changes in the individual behaviour of every group member and reveal the emergent dynamics of turning. First, we show that the turns start from the individuals located at the elongated tips of the flocks. The information to turn then propagates across the whole group with a linear dispersion law and negligible attenuation, therefore minimizing group decoherence. Second, we find that birds on the tips deviate from the mean direction much more persistently than other individuals, indicating that persistent localized fluctuations are a trigger for collective directional switching. Moreover, our analysis reveals two crucial ingredients which enhance the effect of such noise leading to collective changes of state: the non-symmetric nature of interaction between individuals and the presence of heterogeneities in the topology of the network.

2. Methods and data

2.1. Experiments

European starlings are a common site in Rome during winter, where they populate several roosting sites. Shortly before sunset, while returning to their roost, starlings form sharp-boarded flocks that perform highly synchronised manoeuvres while preserving strong coherence. Data on spontaneously initiated turns, without a presence of a predator, were collected at the site of Piazza dei Cinquecento during winter months between 2010 and 2012. The video sequences of turning flocks were acquired using three high–speed cameras IDT-M5 with monochromatic CMOS sensor with resolution 2288×1728 pixels, shooting at 170 fps. In order to obtain the 3D coordinates of individual birds in a flock we used stereophotography, and in particular the trifocal technique with a system of three synchronized cameras positioned at three different points of view [14]. The flocking events took place typically at a distance of 80-130m in front of the cameras. Typical duration of the recorded events is between 1.8 and 12.9s.

2.2. Tracking

The full 3D trajectories of individual birds for the whole duration of the turn are then reconstructed from the video acquisitions. A newly developed 3D tracking algorithm enabled us to retrieve the 3D spatial positions of the same individual through time using computer vision techniques. Details of the algorithm can be found in ref. [10]. The final data set consists of 12 distinct turning events, as reported in Table 1. A typical collective turn is shown in Figure 1a.

3. Results

3.1. Ranking

The reconstructed 3D trajectory of each individual bird i in the flock, at each time step t, is given by its position, $\mathbf{r}_i(t)$, velocity, $\mathbf{v}_i(t)$, and acceleration, $\mathbf{a}_i(t)$, where

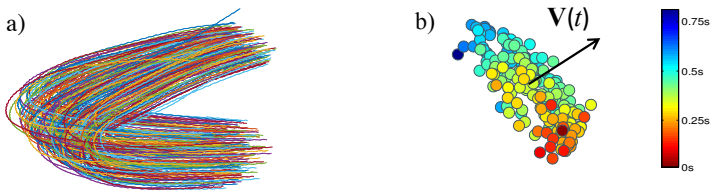


FIG. 1: a) Reconstructed 3D trajectories for a flock of 176 birds performing a collective turn. b) The same flock shown at the start of the turn. Birds are coloured according to their time delays t_i, as indicated by the colour bar, revealing a propagation of the turning information.

the latter two are calculated using a finite difference method (for details see refs. [8, 9]). To discover the dynamics of turning, we establish a ranking of the birds in the flock according to their turning order. We do that by looking at the accelerations of single birds, characterized by a strong peak that signals a turn. By comparing the accelerations curves, we obtain mutual time delays in turning between each pairs of birds in the flock, which in turn enable us to calculate the order and time of turning of each bird. That means that we can tell which bird turned first, which turned second, and so on. Each bird i is then assigned a rank R_i, and its absolute turning time t_i that is equal to the turning delay with respect to the top-ranked bird – the initiator (with delay $t_1 = 0$). Plotting the rank R_i of bird i as a function of its turning time delay t_i gives the ranking curve $R(t)$, see Fig. 2a.

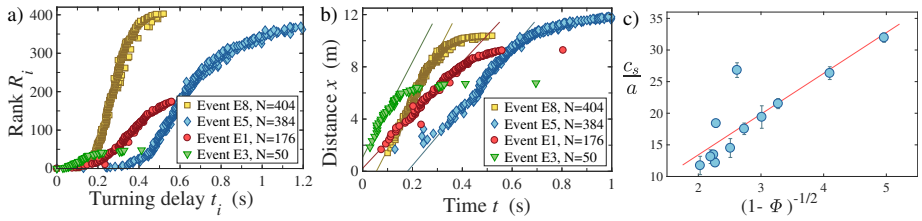


FIG. 2: a) Ranking curves $R(t)$ for several turning events: rank R_i of each bird is plotted vs its turning delay t_i – delay with respect to the first bird to turn. b) Propagation of the turning wave in space: distance x traveled by the information in time t. The speed of propagation, c_s, is the slope of the linear regime of $x(t)$ for early and intermediate times (*full lines*). c) Prediction of the new theory that the rescaled speed of propagation of the turn, c_s/a, must be a linear function of $1/\sqrt{1-\Phi}$, where Φ is the polarization, is verified by the empirical data (P-value: $P = 3.1 \times 10^{-4}$; correlation coefficient: $R^2 = 0.74$). Each point represents a different turning flock. The error bars on c_s are obtained from its variability under changing the linear fitting regime of $x(t)$. The slope of the linear fit (*red line*) is equal to $1/\sqrt{\beta\chi}$ (see equation (7)).

3.2. Start of the turn

The ranking curve gives us several important information about the turn. Its shape indicates that turn is initiated by a small number of birds, whose reaction times are relatively long as the turn starts. If we look at the spatial positions of the birds within the flock in Fig. 1b, we see that the the top-ranked birds (colored in red) are physically close to each other. Moreover, they are located at a tip of a flock, close to one of its lateral sides along the longest elongation axis. Once started, the information to turn then propagates from the initiating birds in all directions, all the way to the birds on the opposite elongated tip of the flock (colored in blue). This means that the decision to turn has a spatially localized origin and it then propagates across the flock through a social transfer of information from bird to bird, as indicated by the spatial modulation of the turning wave.

3.3. Propagation of the turn

Let us first look at the propagation of the directional information travelling through the flock. In ref. [8], the dispersion law was found from the ranking curve $R(t)$ by calculating how much distance x the turning information travels in time t. The turns starts from a localized origin and travels in all directions, therefore, $x(t) = [R(t)/\rho]^{1/3}$ in three dimensions, where ρ is the density of the flock. Figure 2b shows a clear linear regime for early and intermediate times for all turning events, so the distance travelled by the turn grows linearly with time, $x(t) = c_s t$. The parameter c_s is the speed of propagation of the directional information. Interestingly, it varies significantly from flock to flock (see Fig. 2b and Table 1). It turns out that this variability can not be explained by differences in flocks' densities or size, as one might naively expect when thinking of a linear propagation of a sound wave. Indeed, what propagates during the turn are fluctuations of orientation, not of density. Moreover, as shown in ref. [8], acceleration data show that the information to turn propagates across the flock with negligible attenuation. In fact, it is precisely the linear and fast propagation, together with the low damping of the signal, what prevents the loss of flock's cohesion during a turn.

3.4. New theory of flocking

For understanding the theoretical mechanism underlying this phenomena, let us look at the predictions of theoretical descriptions of collective motion [15–20]. The key ingredient of almost all of these theories is the alignment dynamics. This means that each individual tends to keep its direction of motion as close as possible to that of its closest neighbours. In terms of velocities, the alignment dynamics can be written as

$$\mathbf{v}_i(t+1) = \mathbf{v}_i(t) + J \sum_{j \in i} \mathbf{v}_j(t) \ , \tag{1}$$

where the sum extends over all nearest neighbours j of bird i, and the noise (temperature) term, is neglected for simplicity. We assume that the alignment

strength J is large, so that that the system is in a deeply ordered phase, just as natural flocks are [20]. In the following we will study the mathematical consequences of the above alignment dynamics without adding other more realistic terms, as they would not change the effects of the alignment term.

We can write the the above update rule (1) in continuous time limit as

$$\frac{d\mathbf{v}_i}{dt} = -\frac{\partial H}{\partial \mathbf{v}_i}, \qquad H = -J\sum_{\langle ij\rangle} \mathbf{v}_i \cdot \mathbf{v}_j \ , \tag{2}$$

where H is the Hamiltonian of the system. To simplify the analysis, we note the fact that the trajectories of birds during the turn lie on a plane (Fig. 1a and refs. [8, 9]). This allows us to use a 2D order parameter, $\mathbf{v}_i = (v_i^x, v_i^y) = v_0\, e^{i\varphi_i}$, where the phase φ_i is the angle between the direction of motion of bird i and that of the flock (the more general $3d$ case is described in [8, 11]). Now we make an assumption (justified by the data) that v_0 is constant during the turn and same for all the birds. Since flock is a highly ordered system, the velocities $\mathbf{v}_i$ differ little from the average flock's one and $\varphi_i \ll 1$. Hence, we can expand the Hamiltonian and obtain

$$H = \frac{J}{2}\sum_{\langle ij\rangle}(\varphi_i - \varphi_j)^2 = \frac{1}{2}a^2 J \int \frac{d^3x}{a^3}\,[\nabla\varphi(x,t)]^2 \ , \tag{3}$$

in the continuous space limit, where a is the average nearest neighbours distance, while J was rescaled by a term v_0^2. We can now write the equation of motion for the phase φ, using Hamiltonian (3),

$$\frac{\partial\varphi}{\partial t} = -\frac{\delta H}{\delta\varphi} = a^2 J\nabla^2\varphi \tag{4}$$

which is a diffusion equation for the phase φ. This gives dispersion law $\omega \sim ik^2$ for the change of the orientation angle, with the following consequences. It predicts that the information to change direction travels as $x \sim \sqrt{t}$, and that this is an overdamped, non-propagating mode, since the frequency is purely imaginary. This is in complete disagreement with the empirical data on turning flocks presented in the previous sections, where we find the linear propagation with almost no damping.

We find that the standard theoretical description of collective motion, and its prediction described above, have two main problems (see ref. [8, 11] for details): i) missing conservation law – due to the invariance of the Hamiltonian under a uniform rotation of the velocities $\mathbf{v}_i$ ($\varphi_i \to \varphi_i + \delta\varphi$) (meaning that all directions of flight are equivalent), the accompanying conservation law may crucially change the dynamics, thereby affecting the dispersion law; ii) neglected behavioural inertia – the fact that the social force from the neighbours, $F_s = aJ\nabla^2\varphi$, controls directly $\dot{\varphi}$, rather than $\ddot{\varphi}$, gives rise to unreasonable dynamics allowed by Eq. (4), such as a bird performing a U-turn in one time step.

In refs. [8, 11], we show that both problems, missing conservation law and neglected inertia, can be solved by an additional kinetic term in the Hamiltonian, $s_z^2(x,t)/2\chi$. Here s_z is the momentum conjugated to phase φ that generates rotations around the z-axis parameterized by φ, while χ is the generalized moment

of behavioural inertia – resistance of a bird to change its radius of curvature when a social force is exerted by its neighbours [8]. The new Hamiltonian then reads

$$H = \int \frac{d^3x}{a^3} \left\{ \frac{1}{2} a^2 J \ [\nabla \varphi(x,t)]^2 + \frac{s_z^2(x,t)}{2\chi} \right\} . \tag{5}$$

There are strong biological justifications for introducing the additional kinetic term in this particular way. It has previously been observed [21], and recently proved [8, 9, 11, 22], that birds in a flock make an equal-radius turn, rather than parallel-path turn, typical of solid bodies, as it keeps the speed v_0 constant through the flock. This exactly corresponds to the rotation of the velocity vectors parametrized by phase φ as described by the Hamiltonian (5).

Hamiltonian (5), through its canonical equations for the variables (φ, s_z), gives rise to a continuity equation: $\partial s_z/\partial t - \nabla \cdot \mathbf{j}_z = 0$. Here, $\mathbf{j}_z(x,t) = a^2 J \nabla \varphi(x,t)$ is the conserved current which transports the spin s_z in the form of the fluctuations of the phase φ. It is precisely this transport of the intrinsic spin of the bird, s_z, which is equivalent to the inverse radius of curvature of its trajectory (as shown in [8]), that gives rise to the collective turn. Due to conservation of the total spin in the system, when a strong misalignment among the group members forms at some point in the flock, this local excess of curvature (i.e. spin) must be transported away, as it can not be dissipated out locally. This mechanism gives rise to a propagating mode–an undamped spin wave–given by the D'Alembert's equation

$$\frac{\partial^2 \varphi}{\partial t^2} - c_s^2 \ \nabla^2 \varphi = 0 \ , \qquad c_s^2 = a^2 J/\chi \ . \tag{6}$$

This relation describes turning waves with the linear dispersion relation $\omega = c_s k$, that propagate with speed c_s and no damping. In terms of the distance x traveled by the wave in time t, the dispersion relation can be written as $x = c_s t$. This exactly describes the linear and undamped propagation of the directional change across the flock as observed in the experiments.

Interestingly, the new theory for the collective motion described above not only provides an explanation for the linear and undamped propagation observed in turning flocks, but it also makes a prediction that is able to explain the variability of the speed c_s observed in the data (section 3.3). The speed of propagating phase fluctuations depends on the strength of the alignment between the spins, J, as $c_s^2 = a^2 J/\chi$. While we can not experimentally measure the coupling J, we can determine the polarization, $\Phi = |(1/N)\sum_i \mathbf{v}_i/v_i|$, that measures the degree of order (alignment) in the system. It turns out that, for small phase fluctuations, polarization is a function of the coupling J and the noise level β: $\Phi(J,\beta) = 1 - 1/(\beta J)$, where $1/\beta$ is the temperature (see ref.[20]). This means that the speed of information transfer in turning flocks depends on the polarization as

$$c_s = \frac{1}{\sqrt{\beta\chi}} \ \frac{a}{\sqrt{1-\Phi}} \ . \tag{7}$$

Therefore, at fixed noise level β, the speed of propagation of the turn across a flock must be larger the larger the degree of alignment Φ in that flock. In Fig. 2c, we show c_s/a vs. $1/\sqrt{1-\Phi}$ for all analyzed flocks and the data, indeed,

show a clear linear dependence, in agreement with equation (7). Furthermore, relation (7) between polarization and swift collective decision-making might be the reason behind the strong ordering observed in living system.

3.5. Trigger of spontaneous turns

As discussed in section 3.2, the turns we are studying have all started from a spatially localized origin close to the border, before they propagated across the whole flock. Moreover, they initiated without any change in the external environment (no close-by predator, no obvious sound or other perturbation). It has already been shown in experiments (e.g., locusts [6] and fish schools in laboratory tanks [7]) that collective directional switching can be triggered spontaneously from the intrinsic fluctuations in the individual behaviour. Indeed, during starlings aerial display, flocks keep changing their direction of motion even in the absence of predators or obstacles. In ref. [9], by looking at the fluctuations in the individual motion from the global flock's direction prior to the turn, we find that the turn is triggered by the presence of the repeated deviations from the average motion. In fact, the closer a bird is to an elongated tip of the flock, the more persistently it deviates from an average motion. This can be explained by different boundary conditions that individuals at different locations can experience. While birds in the bulk are surrounded by many neighbours and are protected from the external perturbations, the ones at the border, and in particular at the tips of the flock, can move freely towards the empty space around, experiencing an unbalanced social force by the neighbours. These are the factors that can contribute to individual deviations, increasing the probability of a coherent deviation of a few individuals from the common flight direction, which may, in turn, trigger a spontaneous collective swing.

Finally, it is not yet clear the role of the underlying structure of the communication network in these events. In order to understand deeper what makes the initiators of the turn special and whether we can actually predict a turn before it actually occurs, we performed a theoretical study, motivated by our experimental findings, which sheds some light on the origin of frequent spontaneous changes of direction in biological groups [13]. We find that the causes of high sensitivity to fluctuations in these systems, compared to their physical analogues (systems exhibiting ordering phenomena, e.g. ferromagnets), are: the non-symmetric nature of interactions between individuals, and the presence of local heterogeneities in the topology of the network. These crucial ingredients in the living systems we are considering enhance the effect of noise, leading to collective changes of state on finite time-scales and out-of-equilibrium behaviour. In many biological instances of collective behaviour, non-symmetric interactions and network heterogeneities occur naturally, since the individuals coordinate with each other in a non-reciprocal way and the local connectivity is different at the boundary and in the bulk. Analysis presented in ref.[13], therefore, explains why such systems are sensitive to fluctuations that typically build up at the boundary, as observed in experiments (e.g., wild flocks [8] , fish schools [7]).

TABLE 1: Data for the analyzed turning events: number of birds in the flock N; polarization $\Phi = |(1/N)\sum_i \mathbf{v}_i/v_i|$; speed of propagation of the information, c_s, found by fitting the linear regime of the propagation curve, $x(t)$, with the error obtained from its variability under changing the linear fitting time interval of $x(t)$

EVENT	N	Φ	c_s (ms^{-1})
E1	176	0.806	20.20±0.25
E2	125	0.959	42.64±0.97
E3	50	0.866	32.38±1.68
E4	154	0.940	38.46±1.47
E5	384	0.801	22.74±0.71
E6	502	0.841	23.86±2.45

EVENT	N	Φ	c_s (ms^{-1})
E7	139	0.890	37.32±3.42
E8	404	0.854	37.70±1.63
E9	139	0.808	35.40±0.48
E10	197	0.907	27.54±1.01
E11	133	0.793	18.82±1.55
E12	595	0.757	21.96±2.71

References

1. Couzin, I. D. & Krause, J. Self-organization and collective behavior in vertebrates *Adv. Study Behav.* **32**, 1–75 (2003).
2. Sumpter, D.J.T. The principles of collective animal behaviour *Philos. Trans. R. Soc. B* **361** 5–22 (2006).
3. Radakov, D. V. *Schooling and Ecology of Fish* (New York: J. Wiley, 1973).
4. Potts W.K., The chorus-line hypothesis of man oeuvre coordination in avian flocks *Nature* **309**, 344–345 (1984)
5. Tunstrøm, K., Katz, Y., Ioannou, C.C., Huepe, C., Lutz, M.J., and Couzin, I.D. Collective States, Multistability and Transitional Behavior in Schooling Fish *PLoS Comp. Biol.* **9** e1002915 (2013).
6. Buhl J., Sumpter D.J.T., Couzin I.D., Hale J.J., Despland E., Miller E.R., Simpson S.J. From disorder to order in marching locusts *Science* **312**, 1402–1406 (2006).
7. Rosenthal, S.B., Twomey, C. R., Hartnett, A.T., Wu, H.S. and Couzin, I.D.. Revealing the hidden networks of interaction in mobile animal groups allows prediction of complex behavioral contagion. *Proc. Natl. Acad. Sci. USA* **112**, 4690 (2015).
8. Attanasi A., Cavagna A., Del Castello L., Giardina I., Grigera T.S., Jelić A., Melillo S., Poh, O., Shen E., Viale M. Information transfer and behavioural inertia in starling flocks *Nat. Phys.* **10**, 691–696 (2014).
9. Attanasi A., Cavagna A., Del Castello L., Giardina I., Jelić A., Melillo S., Poh, O., Shen E., Viale M. Emergence of collective changes in travel direction of starling flocks from individual birds' fluctuations *J. R. Soc. Int.* **12** (108), 20150319 (2015).
10. Attanasi, A., Cavagna, A., Del Castello, L., Giardina, I., Jelic, A., Melillo, S., Parisi, L., Pellacini, F., Shen, E., Silvestri, E., Viale, M. GReTA – a novel Global and Recursive Tracking Algorithm in three dimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **37** 2451–2463 (2015).
11. Cavagna, A., Del Castello, L., Giardina, I., Grigera, T., Jelic, A., Melillo, S., Mora T, Silvestri E., Viale M. & Walczak, A. M. Flocking and turning: a new model for self-organized collective motion. *J. Stat. Phys.* **158**, 601 (2015)
12. Cavagna, A., Giardina, I., Grigera, T., Jelic, A., Levine, D., Ramaswamy, S. & Viale, M. Silent flocks: constraints on signal propagation across biological groups. *Phys. Rev. Lett.* **114**, 218101 (2015).
13. Cavagna, A., Giardina, I., Jelic, A., Melillo, S., Parisi, L., Silvestri, E., & Viale, M. Nonsymmetric Interactions Trigger Collective Swings in Globally Ordered Systems. *Phys. Rev. Lett.* **118** 138003 (2017).

14. Cavagna, A., Giardina, I., Orlandi, A., Parisi, G., Procaccini, A., Viale, M. & Zdravkovic, V. The starflag handbook on collective animal behaviour: 1. empirical methods *Anim. Behav.* **76**, 217–236 (2008).
15. Toner, J. & Tu, Y. Flocks, herds, and schools: A quantitative theory of flocking. *Phys. Rev. E* **58**, 4828–4858 (1998).
16. Huth, A., & Wissel, C. The Simulation of the Movement of Fish Schools. *J. Theor. Biol.* **156**, 365–385 (1992).
17. Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., & Shochet, O. Novel type of phase transition in a system of self-driven particles. *Phys. Rev. Lett.* **75**, 1226–1229 (1995).
18. Couzin, I. D., Krause, J., James, R., Ruxton, G. D. & Franks, N. R. Collective memory and spatial sorting in animal groups. *J. Theor. Biol.* **218**, 1–11 (2002).
19. Grégoire, G. & Chaté, H. Onset of collective and cohesive motion. *Phys. Rev. Lett.* **92**, 025702 (2004).
20. Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M. & Walczak, A. M. Statistical mechanics for natural flocks of birds. *Proc. Natl. Acad. Sci. USA* **109**, 4786–4791 (2012).
21. Pomeroy, H. & Heppner F. Structure of turning in airborne Rock Dove (*Columba livia*) flocks. *Auk* **109**, 256–267 (1992).
22. Ballerini, M., Cabibbo, N., Candelier, R., Cavagna, A., Cisbani, E., Giardina, I., Orlandi, A., Parisi, G., Procaccini, A., Viale, M. & Zdravkovic, V., Empirical investigation of starling flocks: A benchmark study in collective animal behaviour. *Anim. Behav.* **76,** 201–215 (2008).

Filtering of repeat sequences in genomes

Ana Jelović[1], Miloš Beljanski[2], and Nenad Mitić[3]

[1] Faculty of Transport and Traffic Engineering, Univeristy of Belgrade, 305 Vojvode Stepe, 11000 Belgrade, Serbia
a.jelovic@sf.bg.ac.rs
[2] Institute of General and Physical Chemistry, Studentski trg 12, 11000 Belgrade, Serbia
mbel@matf.bg.ac.rs
[3] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
nenad@matf.bg.ac.rs

Abstract. Finding repeat sequences in nucleic acids and proteins is of great importance in biology. A number of tools are able to efficiently extract these sequences. If we search for repeated sequences in a completely random computer-generated sequence of any meaningful length we will still find a large number of matches. We developed a method for efficiently estimating the probability of a group of found repeated sequences being randomly occurring, and an accompanying program that finds and then filters the found repeated sequences based on the given probability threshold. What makes our method different from existing ones is that we dont group the results by repeat length only but also by number of occurrences. Even short repeated sequences that happen many times may be statistically significant, or longer repeated sequences occurring just a few times may not be. For the large number of repeated sequences that can be found in a genome if the minimal sequence length is relatively low, our method provides a significant gain in performance and quality of results compared to outputting all the found sequences. The method can be applied to both nucleic acids and protein sequences. We have found that, as previously expected, longer repeated sequences mostly have higher probability that they are statistically significant, but also counterintuitively that for some viruses, for example, shorter repeated sequences are more important than the longer ones.

Keywords: repeat sequences, DNA, protein sequences, statistical filtering

1. Introduction

Finding repeat sequences in nucleic acids and proteins is of great importance in biology and a number of tools are able to efficiently extract these sequences, such as Reputer [5], Repex [4], Repseek [3]. Some of these tools work as heuristics while some find all repeat sequences. If a completely random computer-generated sequence of any meaningful length is searched for repeats, a large number of repeats would still be found. Extracting all repeats from a genome

will find a mixture of repeats that are important for its function and organization and randomly occurring repeats that are effectively noise. We developed a method for efficiently estimating the probability of a group of found repeats being randomly occurring, and an accompanying program that finds and then filters the found repeated sequences based on the given probability threshold. What makes our method different from existing ones is that we don't group the results by repeat length only but also by number of occurrences. Even short repeats that happen many times may be statistically significant, or longer repeats occurring just a few times may not be.

1.1. Definitions of repeat types

Repeat sequence is a pair of substrings in the input sequence that satisfies some conditions. Depending on conditions, we have four types of repeat sequences: direct, inverse and both of these kinds can be complementary or non- complementary repeat sequences or we can just call them repeats.
Direct repeat is a pair of two identical substrings on different positions in the input sequence including overlapping substrings.

Example: .....ACTG.....ACTG....

Complementary direct repeat is a pair of two substrings on different positions in the input sequence and every letter in one substring can be obtained by mapping from the letter on the same position in the other substring. (the mapping is defined as a function, for example in DNA the mapping is a-t and c-g)

Example: .....ACTG.....TGAC....

Inverse non-complementary repeat is a pair of substrings in the input sequence so that one substring in the pair can be obtained by reading backwards the other substring in the pair. (they can be on the same position in input sequence so they include words that are palindromes)

Example: .....ACTG.....GTCA....

Inverse complementary repeat is a pair of substrings where one substring can be obtained from the other by reading backwards while using the mapping defined for complements.

Example: .....ACTG.....CAGT....

In literature these repeat types are also known as mirror repeats, everted, inverted repeats etc. As we can see in the first example ACTG is a direct repeat of length four which contains a shorter repeat for example of length two CT, so we define maximal repeats of all four kinds of repeats as a repeat that can not be extended to a longer repeat. In cases of direct non-complementary repeats that means that the letters on the left of the substrings and the letters on the right

of the substring are not the same (similar for complementary direct repeats). In case of inverse non-complementary the letter on the left side of one repeat is not equal to the letter on the right side of the other substring and vice versa (similar for inverse complementary repeats).

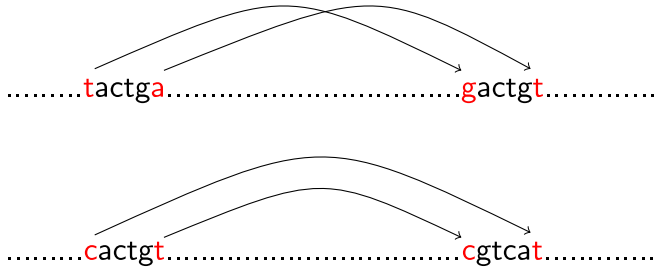


FIG. 1: Letters that should differ in case of direct and inverse non complemenatry repeats are in red colour

2. Methods

Our primary assumption is that a shorter repeat occurring a large number of times can be statistically significant and a longer repeat occurring just a few times may not be. According to this assumption we calculate the expected number of appearance of a repeat length l occurring exactly k times. We calculate this by using different distributions, first calculating the probability of a word of certain length occurring k times. Although the events of a word occurring on different positions in input sequence are not independent events this is done by using Poisson distribution

$$P_s(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!} \tag{1}$$

or Normal distribution [2]

$$P_s(k) = N(\lambda, \lambda - ((2k-1)N - 3k^2 + 4k - 1)\frac{1}{c^2 k}) \tag{2}$$

depending on the word frequency, where

$$\lambda = \frac{N - \ell + 1}{c^\ell}, \tag{3}$$

for N- sequence length, c- alphabet cardinality, ℓ- word length. These formulas hold for non-overlapping words. For overlapping words, whom then tend to occur in clumps Compound Poisson distribution [1] is used

$$P(k) = e^{-\lambda} \sum_{i=1}^{k} \frac{\lambda^i}{i!} \left(\frac{k-1}{i-1} \right) p^{k-i} (1-p)^i \tag{4}$$

or normal distribution [2] if word occurrences are not rare

$$P(k) = N(\lambda, \lambda - ((2k-1)N - 3k^2 + 4k - 1)\frac{1}{c^2 k} + 2\sum_{j=1}^{k-1}(N - 2k + j + 1)\xi_j \frac{1}{c^{2k-j}}) \quad (5)$$

where $\xi_j = 1$. Overlapping words occur rarely comparing to non-overlapping words so it shows that they can be omitted.
Using these calculations we calculate the expected number of repeats with length l occurring k times. First we calculate the expected number of pairs and then multiply it with probability that of all pairs k_l them are maximal repeats.

2.1. Direct non-complementary repeats

In the case of direct repeats non-complementary repeats the pair of substrings that make the repeat are identical. Let $P_{max}(n, k_l)$ denote the probability that among z pairs, k_l of them are maximal. The expected number of lexicographically different repeats (maximal pairs) that occur k_l times for direct non-complementary repeats of length l can be calculated as follows:

$$E_{dn}(k_l) = \sum_{i=n}^{\infty} P_s(i) \cdot P_{max}(i, k_l) \cdot c^l \quad (6)$$

where n is the minimal value that satisfies inequality $k_l \leq \binom{n}{2}$.

2.2. Direct complemenatary repeats

In the case of complementary repeats the pair of substrings that makes the repeat are different substrings.

$$P_{pairs}(z) = \sum_{m \in div(z)} P_s(m) \cdot P_s(\frac{z}{m}) \quad (7)$$

where $div(z)$ represents all divisors of z. If $P(m, \frac{z}{m}, k_l)$ denotes the probability that k_l pairs from z pairs are maximal, then the expected number of lexicographically different repeats (maximal pairs) that occur k_l times for complementary repeats of length l can be calculated using the following formulae:

$$E_{dc}(k_l) = \sum_{i=k_l}^{\infty} \sum_{m \in div(i)} P_s(m) \cdot P_s(\frac{i}{m}) \cdot P(m, \frac{i}{m}, k_l) \cdot c^l \quad (8)$$

2.3. Inverse non-complemenatary repeats

In the case of inverse non-complementary repeats, the pair of substrings that create the repeat are different strings, except in the case of strings that are palindromes themselves. From n substrings that are palindromes themselves $z = \binom{n}{2} + n$ pairs can be made. Let $P(n, k_l)$ denote the probability that from z

pairs, k_l of them are maximal. The expected number of lexicographically different repeats (maximal pairs) that occur k_l times for inverse non-complementary repeats of length l, is

$$E_{pal}(k_l) = \sum_{i=n}^{\infty} P_s(i) \cdot P(i,k_l) \cdot c^{\lceil \frac{l}{2} \rceil} \tag{9}$$

where n is the minimal value that satisfies the inequality $k_l \leq \binom{n}{2} + n$. Adding the expected number for strings that are palindromes themselves to the expected number for other strings, we found that the expected number of lexicographically different repeats (maximal pairs) that occur k_l times for inverse non-complementary repeats of length l is

$$E_{in}(k_l) = E_{dc}(k_l) \cdot (1 - c^{\lceil \frac{l}{2} \rceil - l}) + E_{pal}(k_l) \tag{10}$$

2.4. Inverse complemenatary repeats

In the case of inverse complementary repeats, the pair of substrings that create the repeat are different strings, except in the case of strings that are palindromes themselves. As only strings with even length can be palindromes themselves we use the calculation for inverse non-complementary repeats for repeats with even length and for repeats with odd length we use the calculation for direct complementary repeats.

3. Method implementation

The method implementation has four phases: finding all maximal repeats in an input sequence, calculating the expected number of their occurrence, determining which of the found repeats are statistically significant and outputting the statistically significant repeats. For first phase we used a slightly modified divsufsort algorithm [6], based on suffix arrays. Calculation of the expected number of repeat occurance is performed as previously described. After comparing the calculated confidence bound filtered repeats are outputted in different ways (console, file output or ODBC-compliant database).

3.1. Experimenatl verification of results

As mentioned, all previous calculations are aproximations, so we compared calculated values with repeat counts in randomly generated sequences of same length and cardinality as in input sequence. In Figure 2 we present comparison for results that were found in 1000 randomly generated sequence of length 30000 and calculated expected number of repeats in case of direct complementary repeats length 6. As we can see the the error is relatively low so we can conclude that the approximations are satisfactory. In Figure 3 we present comparison for results that were found in 1000 randomly generated sequence of length 30000 and calculated expected number of repeats in case of direct complementary repeats length 6. Similary the error is relatively low.

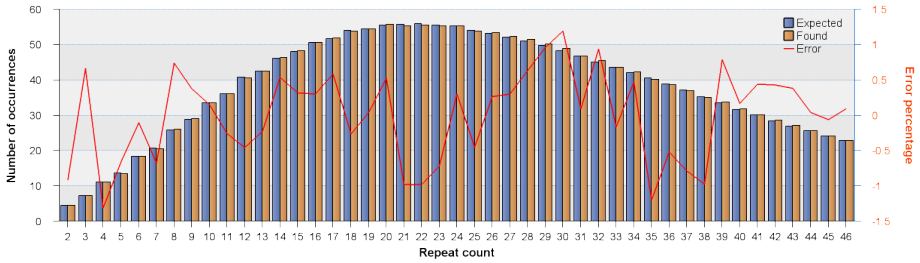


FIG. 2: Results for direct complementary repeats of length 6.The blue bars present the expected number and orange bars present the found number of repeats. On x axes are repeat counts and on left y axis are number of their occurrences. The green line presents the differences between expected and found values in percentage.

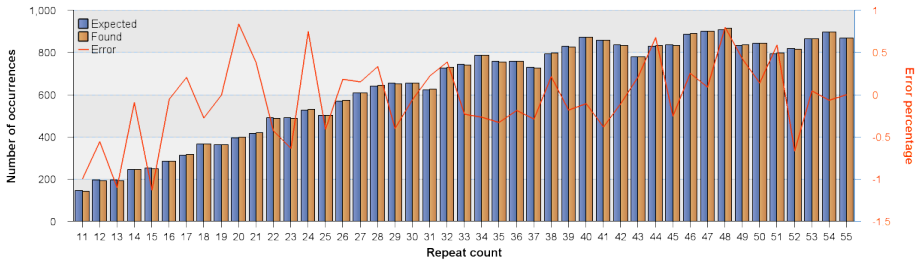


FIG. 3: Comparasion for direct non-complementary repeats of length 8.The orange bars present the found number and blue bars present the expected number of repeats. On x axes are repeat counts and on left y axis are number of their occurrences. The green line presents the differences between expected and found values in percentage.

4. Results

TABLE 1: Statistically sigificant direct complementary repeats for *Zaire Ebolavirus* (AY354458.1)

length	all repeats	statisticaly significant	filtering survival
7	7368	3371	46 %
8	1835	843	46 %
9	499	496	99%
10	107	0	0%
11	23	2	1 %
12	6	0	0 %
13	2	0	0 %
14	1	0	0 %

We used the method on different genomes and found some interesting results. One example is *Zaire ebolavirus* where we searched for direct complementary repeats and the results are shown in table 1. What is interesting is that longer repeats have lower filtering survival then shorter repeats, which is counterintuitive. Starting from some length all found repeats would be statistically significant but this virus does not have repeats that long. Often finding repeats is just the first step in research and sometimes the limit for repeat length is determined arbitrary, expecting that longer repeats are certainly more statistically significant that shorter ones. This example shows that it is not always the case.

On figure 4 we can see results for the same virus, looking just at repeats of length 6. For every expected number we calculate, given the p value, the bound over witch repeat would be statistically significant. P value in this case is 0.05.

TABLE 2: Statistically sigificant direct non-complementary repeats for *Tobacco mosaic* virus (NC_001367.1)

length	all repeats	statisticaly significant	filtering survival
5	13167	5680	43 %
6	3501	1199	34 %
7	926	427	46 %
8	259	259	100 %
9	75	75	100%
10	18	18	100%
11	7	7	100 %
12	1	0	0 %
14	1	1	100 %

Cases like this are not very often. As we expect most of repeats found are statistically significant, which is expected as they for sure mostly present impor-

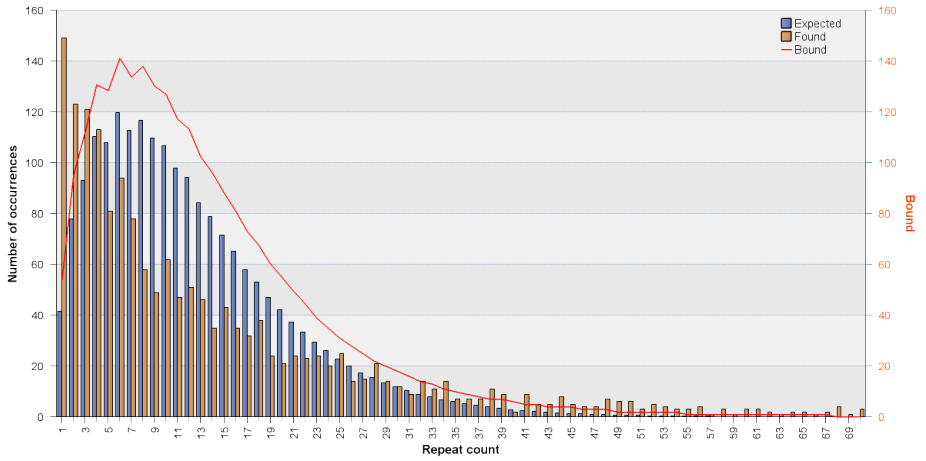


FIG. 4: Comparation of expected and found number of repeats of length 6 for AY354458.1. Orange bars present count of found repeats and blue bars present the expected number that we calculated, that would occur in a randomly generated sequence of same length and alphabet cardinality as the virus itself. The green line presents the bound calculated from expected number, above which repeats are considered statistically significant.

tant signals. In table 2 we se that the majority of longer repeats are statistically significant in case of *Tobacco mosaic*, direct non-complementary repeats.

5. Conclusion

For the large number of repeated sequences that can be found in a genome if the minimal repeat length is relatively low, our method provides a significant gain in performance and quality of results compared to outputting all the found sequences. The method can be applied to both nucleic acids and protein sequences.

Acknowledgments

This work was supported in part by Ministry of Education, Science and Technological Development of the Republic of Serbia, projects No. 174021 and III44006.

References

1. Robin S. and Schbath S. Numerical comparison of several approximations of the word count distribution in random sequences, J Comput Biol.8, 349–359.(2001)
2. Ewens J. W. and Grant R. G. The Analysis of One DNA Sequence in Statistical Methods in Bioinformatics:An Introduction. 2nd ed. Springer Science+Business Media, Inc.(2005)

3. Achaz G., Boyer F., Rocha EP., *et al* Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics. 23, 119–121. (2007)
4. Gurusaran M., Ravella D., Sekar K., RepEx: repeat extractor for biological sequences. Genomics, 102, 403–408. (2013)
5. Kurtz S.,JV. Choudhuri, E. Ohlebusch, *et al*. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl. Acids Re. 29, 4633–4642.(2001)
6. Mori, Y. DivSufSort. (2006)
http://homepage3.nifty.com/wpage/software/libdivsufsort.html.

Could integrative bioinformatic approach predict the circulating miRs that have significant role in pancreatic tissue in type 2 diabetes?

Ivan Jovanović[1], Maja Živković[1], Jasmina Jovanović[2], Tamara Djurić[1], and Aleksandra Stanković[1]

[1] VINČA Institute of Nuclear Sciences, University of Belgrade, Laboratory for Radiobiology and Molecular Genetics, Mike Petrovica Alasa 12-14, 11001 Belgrade, Serbia
{ivanj,majaz,tamariska,alexas}@vin.bg.ac.rs
[2] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia mr06080@matf.bg.ac.rs

Abstract. The action of microRNAs (miRs) as post-transcriptional regulators of gene expression is being recognized as one of the critical processes that affect type 2 diabetes (T2D) progression. The cell–cell signaling via paracrine or even endocrine routes is mediated by miRs released from human tissue. Therefore, the aim of our study was to bioinformatically predict the miRs from microarray gene expression analysis of the whole blood that play role in the pancreas β cell functioning in human T2D. We have demonstrated that gene expression signatures identified in the whole blood correspond to the miR expression changes specific for the pancreas tissue during the insulin resistance. Further experimental studies should follow in order to characterize described effects as early prognostic biomarkers of insulin resistance and T2D.

Keywords: type 2 diabetes, microRNA, microarray gene expression, bioinformatic integrative approach

1. Introduction

Type 2 diabetes (T2D) is a complex disease generally characterized by insulin resistance and increased hepatic glucose production. The rapidly increasing prevalence of T2D is motivating the intensive search for biomarkers of the disease as well as novel therapeutic targets. The action of microRNAs (miRs) as post-transcriptional regulators of gene expression is being recognized as one of the critical processes that affect T2D progression. Therefore, these small, non-coding RNAs, that regulate gene expression by predominantly promoting the degradation of mRNA [1], exhibit great biomarker and therapeutic potential [2]. Also, it was described that all human cells can release miRs, which mediate cell–cell signaling via paracrine or even endocrine routes [3].

Recently, microarray whole genome expression data and miR target predictions from multiple prediction algorithms was linked using a multivariate statistical

technique called Co-Inertia analysis (CIA) in order to predict miR activity and to associate specific miRs with different diseases [4–6]. The studies have shown that CIA method does provide good quality predictions of miR activity [4–6]. It was suggested that CIA has complementarity with other previously described prediction approaches [7] thus could offer the prediction of miRs unidentified by others. So far, this integrative approach was not used for the analysis of circulating miRs that may originate from pancreatic tissue in T2D. Therefore, the aim of our study was to bioinformatically predict the miRs from microarray gene expression analysis of the whole blood that play role in the pancreas β cell functioning in human T2D.

2. Materials and Methods

2.1. Gene expression data

The gene expression data set used in our study was downloaded from http://www.ncbi.nlm.nih.gov/geo/ (Gene Expression Omnibus database), accession number: GSE26168 [8]. The total mRNA expression of whole blood from T2D patients and healthy controls was profiled on Illumina HumanRef-8 v3.0 expression beadchip. The data was initially background subtracted and quantile normalization was performed prior the analysis.

2.2. Co-Inertia analysis

CIA was used to link microarray gene expression data (8 T2D patients and 8 controls) and miR target predictions from multiple prediction algorithms to associate specific miR activity with T2D. This multivariate statistical technique simultaneously analyzes two connected data tables. The tables are treated as two sets of measurements on the same objects, genes. One of the tables is the mRNA gene expression table of g genes from n samples and the other displays predicted target counts of all miRs for the same g genes. Non-symmetric correspondence analysis was used as ordination method of CIA, which summarizes each data table in a low dimensional space by projecting the samples onto axes which maximize the variances of the coordinates of the projected points. CIA performs two simultaneous NSCs on the two linked tables, and identifies pairs of axes, from the two datasets which are maximally covariant. This unsupervised method was used for visual inspection of the data only. By further use of Between Group Analysis (BGA) [9], which forces an ordination to be carried out on groups of samples rather than individual samples, CIA was directed to find the maximum co-variance between the gene expression difference between groups of samples and the miR-gene target frequency tables [4]. For the specified split in the data that contrasts T2D and control samples, we received a ranked list of miR motifs. CIA generates as many miR rank lists as target prediction algorithms used. The most extreme values of the ranking lists (top 20 and last 20) were used for the prediction of upregulated and downregulated miRs in T2D. Lists were combined using consistency among the methods, according to previous study [4]. The complete analysis was performed by the MADE4 R package [10].

2.3. miR target prediction

Five sequence based miR target prediction algorithms were used for CIA: TargetScan and TargetScanS, PicTar4way and Pictar5way, and miRanda according to Madden et al [4]. Each of these sequence based prediction algorithms utilizes the complementarity of mRNA target site with the miR seed and the cross species conservation in their predictions. The miR target prediction data for CIA input, extracted from these databases, was organized in gene/miR frequency tables of counts of predicted targets per gene for each of the algorithm. The gene/miR frequency tables for sequence based predictions originated from the TargetScan website http://www.targetscan.org/ (version 4.1), the UCSC genome browser tract for pictar4way and pictar5way http://genome.ucsc.edu/, and from miRBase for miRanda (http://microrna.sanger.ac.uk/sequences/). The gene/miR frequency tables were provided by the authors [4].

3. Results

CIA was firstly used in unsupervised manner for the purpose of data exploration. Figure 1 shows an example of unsupervised analysis of CIA using Pictar4Way target prediction program. The plot is in 2 parts and depicts a correspondence analysis of T2D patients and control samples and miRs associated with the gene expression pattern characteristic for the two groups of samples. The observed split in the data shows clear difference between the gene expression profiles of the analyzed groups (Figure 1). The CIA performed in conjunction with correspondence analysis and between group analysis produced five ranked lists of miRs associated with specific gene expression profile in T2D. Using consistency among methods, we characterized potentially upregulated and downregulated circulating miRs responsible for the whole blood gene expression template in T2D (data not shown).

Clear clustering of T2D samples and controls shown in Fig. 1. depict homogeneous genome expression from whole blood (from microarray experiment) in T2D patients and different from control samples. This makes data suitable for further, supervised CIA. Our preliminary results indicate successful prediction of miRs from blood and applicability of our approach to select T2D associated miRs, as potentially new molecular biomarkers for this disease.

By inspecting the supervised CIA results, along with literature mining, we discovered that two of the highly ranked miRs (Table 1) present important factors in pancreas β cell proliferation in response to hyperglycemia and insulin resistance which is the hallmark of T2D.

4. Discussion and conclusion

Blood miRs expression patterns have been reported for various human diseases with disease specific signatures [11]. In one of the first studies, it was shown by sequencing that patients with T2D have a significantly altered expression profile of serum miRs [12]. This approach was also favored in the detection of miRs in

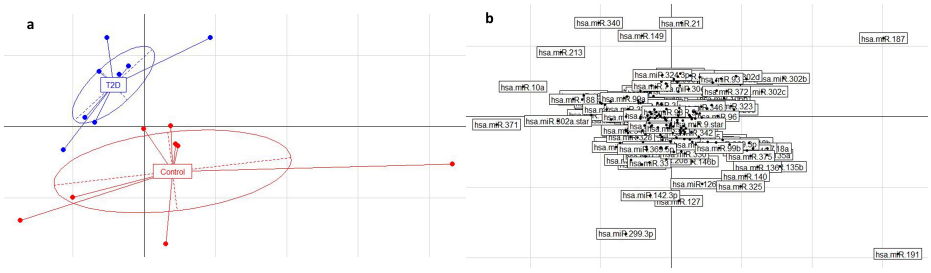


FIG. 1: Axes of the unsupervised CIA performed on the whole genome gene expression data of T2D patients and controls. The gene/miR frequency table generated with Pictar4Way was used to make this figure. a) shows the projection of T2D and control samples, b) shows the projection of the miRs associated with different groups of samples.

TABLE 1: The ranking of the selected miRs according to CIA performed on 5 prediction algorithms

miR	*P4W*	*P5W*	*TS*	*TSS*	*Miranda*	*Predicted regulation in T2D*
miR-375	6	-	11	2	-	UP
miR-184	8	1	17	7	-	DOWN

P4W Pictar4Way; P5W Pictar5Way; TS TargetScan; TSS TargetScanS

blood and other body fluids of T2D patients [8, 13–15].

Using bioinformatical approach that combines microarray gene expression and miR target prediction from multiple prediction algorithms, we have associated specific circulating miRs with T2D. In this study, we have focused on the two of the most noteworthy miRs, functionally associated in a network within the miR pathway that coordinately regulates the compensatory proliferation of the pancreatic β cells in T2D.

The miR-184 is unique in pancreatic islets as the most downregulated miR during insulin resistance [16]. It was described that miR-184 acts as an inhibitor of Ago2 [16]. Increased expression of Ago2 facilitates the function of already upregulated miR-375 in suppressing genes, including growth suppressor Cadm1 in vivo, thus inducing the proliferation of β cells and accommodation of the elevated demand for insulin [16]. Therefore, the miR-184 mir-375 network presents the essential component of the compensatory response that regulates proliferation of β cells regarding insulin sensitivity and metabolic stress [16].

The most important finding of our study is that the whole blood gene expression signatures reflects the miR expression changes specific for the pancreas tissue during the insulin resistance. This is the first bioinformatical study showing that tissue-released miRs affect the whole blood gene expression in T2D. Although there is still a debate about the hormone-like effect of extracellular miR in the blood [3], the results of our study suggest that certain circulating miRs could be systemic biomarkers of pancreatic tissue changes in T2D. The results of our pre-

dictions are also in agreement with microarray expression results of circulating miRs in T2D [8].

The obtained results represent the data of great importance for understanding of complexity of miR nature. Also, here we demonstrate the crucial need of bioinformatical integrative concepts in further research of molecular processes of T2D. Finally, further experimental studies should follow in order to characterize described effects as early prognostic biomarkers of insulin resistance and T2D.

Acknowledgments

This work was supported by Serbian Ministry of Education, Science and Technological development Grant OI175085.

References

1. Guo, H., Ingolia, NT., Weissman, JS., Bartel, DP.: Mammalian microRNAs predominantly act to decrease target mRNA levels. Nature, Vol. 466, No. 7308, 835-40. (2010)
2. Chen, K., Rajewsky, N.: The evolution of gene regulation by transcription factors and microRNAs. Nature Reviews Genetics, Vol. 8, No. 2, 93-103. (2007)
3. Turchinovich, A., Samatov, TR., Tonevitsky, AG., Burwinkel, B.: Circulating miRNAs: cell-cell communication function? Frontiers in Genetics, Jun 28;4:119. (2013)
4. Madden, SF., Carpenter, SB., Jeffery, IB., Bjrkbacka, H., Fitzgerald, KA., O'Neill, LA., Higgins, DG.: Detecting microRNA activity from gene expression data. BMC Bioinformatics, 11: 257. (2010)
5. Mulrane, L., Madden, SF., Brennan, DJ., Greme, G., McGee, SF., McNally, S., et al, DP.: miR-187 is an independent prognostic factor in breast cancer and confers increased invasive potential in vitro. Clinical Cancer Research, Vol. 18, No. 24, 6702-13. (2012)
6. Jovanovi, I., Zivkovi, M., Jovanovi, J., Djuri, T., Stankovi, A.: The co-inertia approach in identification of specific microRNA in early and advanced atherosclerosis plaque. Medical Hypotheses, Vol. 83, No. 1, 11-5. (2014)
7. Arora, A., Simpson, DA., Individual mRNA expression profiles reveal the effects of specific microRNAs. Genome Biology, Vol. 9, No. 5, R82. (2008)
8. Karolina, DS., Armugam, A., Tavintharan, S., Wong, MT., Lim, SC., Sum, CF., Jeyaseelan, K.: MicroRNA 144 impairs insulin signaling by inhibiting the expression of insulin receptor substrate 1 in type 2 diabetes mellitus. PLoS One, Vol. 6, No. 8, e22839. (2011)
9. Culhane, AC., Perrire, G., Considine, EC., Cotter, TG., Higgins, DG.: Between-group analysis of microarray data. Bioinformatics, Vol. 18, No. 12, 1600-8. (2002)
10. Culhane, AC., Thioulouse, J., Perrire, G., Higgins, DG.: MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics, Vol. 21, No. 11, 2789-90. (2005)
11. Keller, A., Leidinger, P., Vogel, B., Backes, C., ElSharawy, A., Galata, V., et al.: miRNAs can be generally associated with human pathologies as exemplified for miR-144. BMC Medicine, 12:224. (2014)
12. Chen, X., Ba, Y., Ma, L., Cai, X., Yin, Y., Wang, K., et al.: Characterization of microRNAs in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Research. Vol. 18, No. 10, 997-1006. (2008)
13. Collares, CV., Evangelista, AF., Xavier, DJ., Rassi, DM., Arns, T., Foss-Freitas, MC, et al. Identifying common and specific microRNAs expressed in peripheral blood mononuclear cell of type 1, type 2, and gestational diabetes mellitus patients. BMC research notes, 6:491. (2013)

14. Zampetaki, A., Kiechl, S., Drozdov, I., Willeit, P., Mayr, U., Prokopi, M., et al.: Plasma microRNA profiling reveals loss of endothelial miR-126 and other microRNAs in type 2 diabetes. Circulation Research, Vol. 107, No. 6, 810-7. (2010)
15. Kong, L., Zhu, J., Han, W., Jiang, X., Xu, M., Zhao, Y., et al.: Significance of serum microRNAs in pre-diabetes and newly diagnosed type 2 diabetes: a clinical study. Acta Diabetologica, Vol. 48, No, 1, 61-9. (2011)
16. Tattikota, SG., Rathjen, T., McAnulty, SJ., Wessels, HH., Akerman, I., van de Bunt, M., et al.: Argonaute2 Mediates Compensatory Expansion of the Pancreatic Cell. Cell Metabolism, Vol. 19, No. 1, 122-134. (2014)

A biologically-inspired model of visual word recognition

Yair Lakretz[1], Naama Friedmann[1], and Alessandro Treves[2]

[1] Tel-Aviv University, Tel-Aviv, Israel
[2] SISSA, 265 Via Bonomea, Trieste, Italy

Abstract. We present a computational model of visual word recognition. The model is biologically inspired, incorporating plausible cortical dynamics, thus adding to previous studies, which have used connectionist or 'box-and-arrow' type models. We begin by exploring several methods to represent the letter identities in an artificial neural network, and to identify the method that best agrees with experimental findings and computational constraints. In the self-organization process of a multilayer neural network, letter-identity and letter-position representations are further processed to create word representations. These correspond to word memories in an orthographic lexicon, as described in neuropsychological models, and function as attractors of the neural network. Simulations present normal reading by the network in the absence of noise or deficits. When noise or deficits are introduced, the network presents failures such as letter transposition or letter substitution, which are similar to those made by dyslexics with letter-position dyslexia and letter-identity dyslexia, respectively.

Keywords: Reading, attractor neural networks, dyslexia

1. Introduction

Reading is a complex skill. It requires the brain to perform multiple processes such as graphical pattern recognition, extraction of meaning, word production and more, all in parallel and in strikingly short time. The first stages of the process of reading include the encoding of letter identities, letter position processing, and the composition of letters into words. Neuropsychological studies have shown that these functions can be selectively impaired and give rise to specific dyslexias [7]. Most importantly for the current study, a dyslexia has been identified in which letter position encoding is impaired [3, 4, 6]. Several computational models for visual word recognition (VWR) have been proposed in the literature [8, for a review]. Although insightful and comprehensive, these models shed little light on how the brain performs these tasks. We hereby present a model that brings cognitive models together with plausible brain dynamics. These are modeled in an attractor dynamics network consisting of graded-response neurons with threshold-linear activation function [9]. The model addresses the question of how these processes are executed at the neuronal level, including possible failures in processing, due to noise or deficits.

2. The model

2.1. Letter representations

This section investigates the very first stage of reading, from the printed word to the level of letter representation. The activation created by a printed-letter input in an early visual stage is modeled by ascribing a list of factors to a letter from all possible graphical features in a letter (figure 1A). Factors create in turn activation in a higher layer, which we name *the letter layer*. Letter representations are then used to compose written words at the orthographic lexicon.

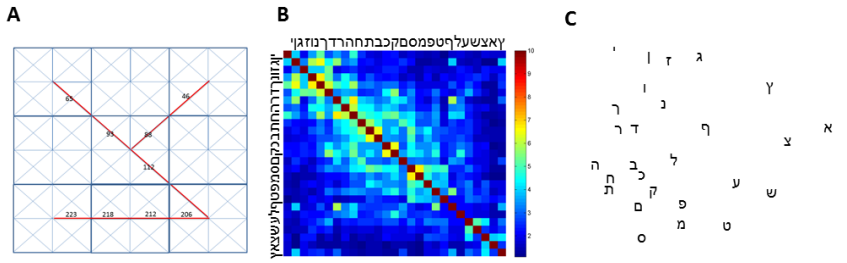


FIG. 1: (A) Example of visual factors in the representation of a Hebrew letter. (B) Letter similarity among all letter pairs in Hebrew as judged by Hebrew readers. Scores are between 1 and 10. (C) Multidimensional scaling of all 27 Hebrew letters.

In order to reduce interference in memory retrieval between words in the lexicon, letter representations should be as little correlated as possible. We examine and compare several methods for the generation of letter representation in this early stage of reading. All methods are taken to be simple abstractions of possible neuronal processes in the brain:

Constituting factors This method assumes a two-layer architecture: a factor- and a letter-representation layer. Each feature in the graphical form of a letter is represented in the model as a unit in the factor layer. Letters that have same features will hence share the corresponding active-units. Each unit in the factor layer creates in turn activation in a predefined random subset of units in the letter layer.

Renormalization As in the first method, each printed-letter input creates activation in a factor layer. In this case however, the contribution of each factor to the final representation is increased by a factor that is inversely proportional to its appearance in other letters. Salient features will therefore have higher weight in the final letter representation. In addition, a competition between neighboring features occurs, leaving in that neighborhood only features that are most salient.

Intermediate sub-network layer A third layer between the factor and letter layers is added. This layer is composed of several sub-networks; each sub-network corresponds to a receptive-field (RF), which is a surrounding of neighboring cells. The size of the RF is a parameter of the model. The optimal value of this parameter will be investigated below. Each factor is connected to a random subset

of units in the sub-network layer. The size of this subset (UPF units per factor) is another parameter of the model later to be determined. The connections between the sub-network layer and letter representation layer are set in the following way: each weight between a unit in the sub-network layer and the representation layer is inversely proportional to the accumulated activation in that sub-network unit across all letters. That is, popular units in the sub-network are less dominant in the final pattern of activation. Therefore, similar to the second method, salient features have higher weight in the letter representation than features with high occurrence.

Figure 2 presents correlation matrices for the 27 letters in the Hebrew alphabet for the three methods. For each correlation matrix (top), a corresponding full-cue retrieval test is presented along with (bottom). A full-cue test is done by presenting the network with a full-cue of the printed letter and counting the number of times successful retrieval occurs. Results show that low values of the correlations matrix correspond to high full-cue retrieval performance, and that the intermediate sub-network layer method achieves best performance. We therefore focus on this method in what follows.

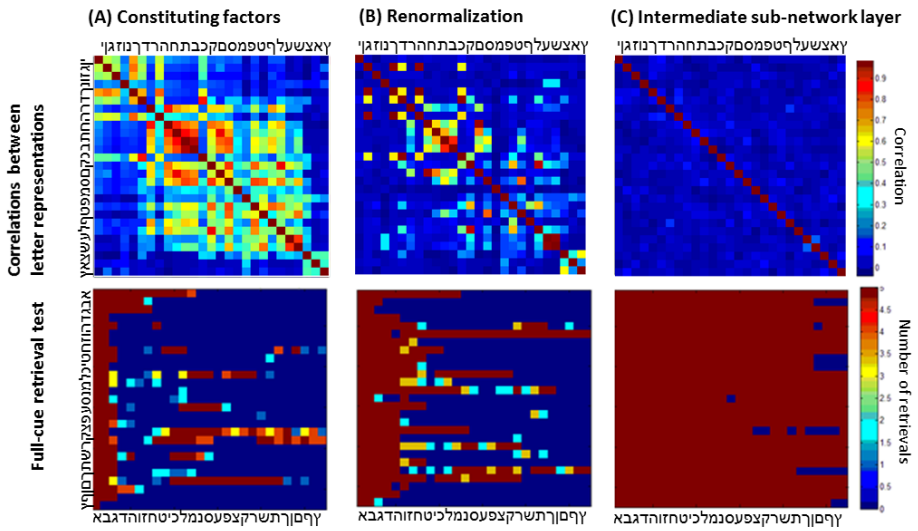


FIG. 2: Correlation matrices and full-cue retrieval test results for the three methods. (A) Constituting factors (B) Renormalization (C) Sub-network Layer (RF=1, UPF=100).

Note, however, that in addition to low correlations between letter representations, we require that similarity between these representations will correspond to letter similarity as judged by readers. That is, taking in consideration both these constraints, letter representations cannot be completely orthogonal. Figure 1B presents average similarities between letters as judged by 30 subjects. In this test, subjects were asked to judge similarities among all letter pairs in the Hebrew alphabet. We use this data to determine the optimal model parameters

by choosing the values that: (a) maximize the correlation between letter similarities according to the model and the experimental data; and (b) minimize the mean correlation between letter representations of the model.

2.2. Word representations and the learning stage of the network

A full description of the composition of words from letters is beyond the scope of this report. The process of composition is done in two steps. First, for each serial letter, letter identity and position are composed together. This is done through a competition process between letter-identity and letter-position activations. Next, all resulting letter-in-their-position representations are composed together. This is done by a similar competitive process, eventually creating the desired word-representation. Importantly, units that encode letter position are the same units that encode letter identity, which is presumably the case in the brain.

The final word representations undergo a follow-up self-organization process, which reduces redundant correlations between the representations. In this process, a multilayer neural network, endowed with Hebbian learning and synaptic scaling, is repeatedly presented with word patterns in a random order. The resulting word representations are finally stored in the final layer of the network, which we name *the word layer*, and function as its attractor states.

2.3. Architecture and dynamics

The complete architecture of the model is a multilayer network, starting at the factor layer and ending at the word layer. Units in the network are graded-response neurons, that is, a positive continuous variable V_i that is proportional to the activity of the neuron is assigned to every unit. This is in accordance with an interpretation of V_i as mean firing-rates.

The updating of the network assumes a threshold-linear activation function: $V(t) = g(h(t) - \theta)\theta(h(t) - \theta)$, where $h(t)$ is the local field, which in the word layer amounts to summation over all excitatory inputs: $h_i = V_{input} + \sum_j W_{ij} V_j$; θ is the threshold below which there is no output; g is a gain factor; and W_{ij} are the synaptic weights as defined below. Each update step is followed by a competitive process which brings the sparseness of the network to a constant value. The sparseness a is defined as: $a = \frac{(\sum_i V_i/N)^2}{\sum_i V_i^2/N}$, which in the limit of the binary case is equivalent to the fraction of active units. We set this value to $a = 0.25$, which is in the range of plausible cortical values. This competitive process represents inhibitory feedback regulation on the activation of the network, and it operates by adjusting the threshold and gain parameters of the threshold-linear activation function.

Connections between neurons at the word layer are according to a covariance Hebbian rule: $W_{ij} = \frac{1}{a} \sum_\mu \xi_i^\mu (\xi_j^\mu - \bar{\xi})$, where ξ^μ is the μ'th word pattern, and $\bar{\xi}$ is the mean across all words.

After the learning stage described above is over, new words can be presented to the network. Activations created by the printed word flow from the factor layer, in a feed forward manner, to the word layer, finally converging according to the above dynamics. The resulting pattern can then be compared to the stored memory patterns in the lexicon.

Since similarities among letter identities and among letter positions are incorporated in the model, the network exhibits several phenomena: under noise conditions, a printed-word input can cause the network to converge to an incorrect attractor, which corresponds to the printed-word with transposed positions of letters, or to a word in which one letter is replaced by a similar one. These phenomena are dependent on the amount of noise and deficits presented to the network, and correspond to errors as described in dyslexia [3, 1, 5]. Further work is required to compare error statistics of the network to those found in reading test results.

3. Summary

We have presented a biologically-inspired computational model of visual word recognition. We have explored several methods for the representations of letter identity. The method that achieved best performance was that of adding an intermediate layer with sub-networks, which correspond to visual receptive fields. The optimal parameter values of this method were determined by two constraints: (a) low correlations between letter representations (to improve memory capacity of the neural network); and (b) high correlation between similarity relations among letter representations in the model, and those found in behavioral tests. Simulations in the absence of noise or deficits show almost perfect retrieval of word memories. When noise or deficits are presented, the network exhibits reading errors such as letter transposition or substitution, similarly to dyslexics. A full report of the results of the simulations will be presented elsewhere.

References

1. Brunsdon, Ruth and Coltheart, Max and Nickels, Lyndsey: Severe developmental letter-processing impairment: A treatment case study. Cognitive neuropsychology, 2006, Vol. 23, No.6, pp.795–821 (2006)
2. Coltheart, Max and Rastle, Kathleen and Perry, Conrad and Langdon, Robyn and Ziegler, Johannes: DRC: a dual route cascaded model of visual word recognition and reading aloud, Psychological review, Vol. 108, No. 1, pp. 204 (2001)
3. Friedmann, Naama and Gvion, Aviah: Letter position dyslexia, Cognitive Neuropsychology, Vol. 18, No. 8, pp.673–696 (2001)
4. Friedmann, Naama and Rahamim, Einav: Developmental letter position dyslexia, Journal of Neuropsychology, Vol. 1, No. 2, pp. 201–236 (2007)
5. Friedmann, Naama and Biran, Michal and Gvion, Aviah: Patterns of visual dyslexia, Journal of neuropsychology, Vol. 6, No. 1, 1–30 (2012)
6. Friedmann, Naama and Rahamim, Einav: What can reduce letter migrations in letter position dyslexia?, Journal of Research in Reading, Vol. 37, No. 3, pp. 297–315 (2014)
7. Friedmann, Naama and Coltheart, Max and Bar-On, A and Ravid, D: Types of developmental dyslexia. In Handbook of communication disorders: Theoretical, empirical, and applied linguistics perspectives, eds A. Bar-On and D. Ravid (De Gruyter Mouton)
8. Norris, Dennis: Models of visual word recognition, Trends in cognitive sciences, Bol. 37, No. 10, pp.517–524 (2013)
9. Treves, Alessandro: Graded-response neurons and information encodings in autoassociative memories, Physical Review A, Vol. 42, No. 4, pp. 2418 (1990)

A Quantum Approach to the DNA Functioning

A. Nicolaidis

Department of Theoretical Physics,
Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
nicolaid@auth.gr

Abstract. In the present work we prime the notion that DNA is an information processing system, receiving registering transferring information. In the pursuit of an inherent logic in DNA functioning, we explore the possibility that quantum logic might serve this purpose. We use the quantum formalism to describe the DNA dynamics and as a byproduct we obtain the DNA vacuum. The DNA vacuum, in clear analogy to the quantum vacuum, is a collection of virtual DNA bases. An essential aspect of the DNA functioning is the complementarity relation R, which binds the pairs AT, GC, and generates the replication process. Further in an effort to codify DNA, we introduce Gdels numbering for a DNA strand, assigning a specific natural number to each individual strand. This numbering allows a quantitative measure of the difference among the various DNA strands. Considering also that the four DNA bases constitute an "alphabet", we may assume the task to examine if DNA is a "language".

1. Introduction

Considering the entire evolution, from cosmology, through biology, to language, we gather that biology verges on natural sciences and linguistics. Natural sciences (physics, chemistry) prime the notion of structure. There are elementary constituents, the "atoms". The "atoms" through interactions, get composed to give larger structures (from quarks, to protons and neutrons, nuclei, atoms, molecules, stars, galaxies). Linguistics primes the notion of information. It introduces the "sign", or the "word", which denotes and refers to another object.

We expect that biology shares features from both forms of knowledge, natural sciences and linguistics. Indeed in biology we encounter the biological atoms, the four nucleotide molecules (adenine, guanine, cytosine and thymine). Further the four nucleotides get composed to form larger structures, the DNA sequences, amino acids, proteins. From another point of view the four nucleotides may be considered as not simply the constituents of biological structures, but as the "letters" of a language. These "letters" give rise to biological "words", "phrases", "sentences". The biological "words" or "phrases" act like signs, receiving registering transferring information, executing specific functions, favoring or disfavoring a biological process. It is an open and a highly interesting question if the biological "text" follows an internal logic, or a syntax.

In the present work, in the search of a biological syntax, we would like to explore the possibility that quantum logic might be related to the internal logical functioning of DNA. Most of the scientific edifice relies on Aristotles logic

(with the law of excluded third: it is either A or its opposite ¬ A). Quantum Mechanics though, defies common sense and common logic. In ref. [1, 2] we have suggested that the relational logic of C. S. Peirce [3, 4] may serve as the conceptual foundation of quantum mechanics and string theory. Within relational logic, relation is the primary irreducible datum and everything is expressed in terms of relations. We are led to reorient our thinking and consider that things have no meaning in themselves, and that only the correlations between them are "real". Few examples of relations R_{ij} might be indicative: the transition from a state **j** to a state **i**, the proof of a theorem **i** starting from a theorem **j**, the transformation of a metabolite j to a metabolite i. Relations may be composed and a third transitive relation emerges following the rule

$$R_{ij}R_{kl} = \delta_{jk}R_{il} \tag{1}$$

In quantum theory, the quantum states are living in a Hilbert space. Transitions between the quantum states are realized through relations or projection operators. We propose then to represent the four nucleotides (A, G, C and T) as states living in an abstract space. The pairings AT and GC are achieved through a corresponding projection operator. Furthermore, we introduce the concept of DNA vacuum. In a similar way that the quantum vacuum is not empty, but is a collection of virtual particles, the DNA vacuum is a collection of all possible bases. The notion of the DNA vacuum will appear very useful in better understanding the replication process. In the next section we introduce the quantum formalism and we use it in order to represent and better understand the DNA functioning. In the third section we consider a DNA strand as a "theorem" in logic. Following Gdels numbering of mathematical theorems [5], we represent the DNA strand by a unique natural number, a product of prime numbers. This encoding of DNA may allow a quantitative measure of the difference among the various DNA strands. Also it would help for a faster and more efficient analysis of DNA through data mining techniques. In the last section we present our conclusions and indicate directions for future research.

2. Quantum Formalism for DNA

We suggest that the quantum formalism is apt in order to describe the inherent DNA dynamics. Let us remind the essentials of the quantum formalism, using the Dirac braket notation. A quantum ket state $|S_i\rangle$ is living in an abstract Hilbert vector space. Next to the Hilbert space there is a dual Hilbert space where live the bra states $\langle S_i|$. A bra state is the transpose and conjugate of the corresponding ket state. We can define then two products. The "inner product"

$$A_{ij} = \langle S_i | S_j \rangle \tag{2}$$

is a number and indicates the affinity or similarity between the $|S_i\rangle$ and $|S_j\rangle$. Since the "inner product" is in general a complex number, a real measure of the affinity is obtained through the probability p ($0 \leq p \leq 1$)

$$p = A_{ij}A_{ji} \tag{3}$$

Like in all vector spaces, in our Hilbert vector space there are "base vectors" which are orthonormal, satisfying

$$\langle n|m\rangle = \delta_{nm} \tag{4}$$

We define also the "outer product"

$$R_{ij} = |S_i\rangle \langle S_j| \tag{5}$$

R_{ij} stands for a relation expressing the transition from the initial state $|S_j\rangle$ to the final state $|S_i\rangle$.It can be considered also as a projection operator which allows as incoming state $|S_j\rangle$ and as outgoing state $|S_i\rangle$. Clearly R_{ij} , thus defined, satisfies the composition rule eq. (1).

The hereditary information of an organism is encoded in the DNA. The DNA is a macromolecule composed of two polynucleotide chains with a double-helical structure. There are four distinct bases, building elements of the genetic information: adenine, cytosine, guanine, thymine, abbreviated A, C, G, T respectively. A single strand DNA is simply a chain of nucleotides where two consecutive nucleotides are bound together by a strong covalent bond. Each single strand has a natural orientation. This orientation is due to the fact that one end of the single strand has a free 5' phosphate group and the other has a free 3' deoxyribose hydroxyl group. The most important feature of DNA is the Watson-Crick complementarity of bases. Bonding between single strands occurs by the pairwise attraction of bases: A bonds with T and G bonds with C. The pairs (A, T) and (G, C) are therefore known as complementary base pairs. The classical double helix of DNA is formed when two separate stands bond. Two requirements must be met for this to occur; firstly, the strands must be complementary, and secondly, they must have opposite orientations.

We may now borrow quantum ideas and apply them to the DNA functioning. We suggest that the four bases A, C, G, T, are states belonging to a Hilbert space and we represent them by $|B_1\rangle$, $|B_2\rangle$, $|B_3\rangle$ and $|B_4\rangle$ respectively (clearly the number correspondence is arbitrary). The states satisfy the orthonormal condition

$$\langle B_i|B_j\rangle = \delta_{ij} \tag{6}$$

The bonding between the bases $|B_1\rangle$ and $|B_4\rangle$ is achieved through a relation (or equivalently a projection operator) P

$$P = |B_1\rangle \langle B_4| + |B_4\rangle \langle B_1| \tag{7}$$

The bonding between the bases $|B_2\rangle$ and $|B_3\rangle$ is achieved in a similar fashion through a relation Q

$$Q = |B_2\rangle \langle B_3| + |B_2\rangle \langle B_3| \tag{8}$$

The full complementarity is realized with the relation

$$R = P + Q \tag{9}$$

Notice that

$$R|B_i\rangle = |B_j\rangle \tag{10}$$

where $|B_i\rangle$ and $\left|B_j\right\rangle$ are complementary pairs. R acts like a "mirror" operation with A, T being the images of each other (similarly for the C and G bases). R^2 is an idempotent operator, i.e. it is the unity operator $\mathbf{1}$

$$R^2 = \sum_i |B_i\rangle \langle B_i| = \mathbf{1} \tag{11}$$

and

$$R^2 \left|B_j\right\rangle = \left|B_j\right\rangle \tag{12}$$

Let us concentrate on the 2-state problem specified by the P relation (similarly for the Q relation). $|B_1\rangle$ and $|B_4\rangle$ are represented by the columns

$$|B_1\rangle =, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \ |B_4\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{13}$$

and the P relation is represented by the matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{14}$$

The eigenstates of P are the states $|B_+\rangle$ and $|B_-\rangle$

$$|B_+\rangle = \frac{1}{\sqrt{2}} \left[|B_1\rangle + |B_4\rangle\right] \tag{15}$$

$$|B_+\rangle = \frac{1}{\sqrt{2}} \left[|B_1\rangle + |B_4\rangle\right] \tag{16}$$

Notice that the states $|B_+\rangle$ and $|B_-\rangle$ are transformed to each other through the relation S

$$S|B_\pm\rangle = |B_\mp\rangle \tag{17}$$

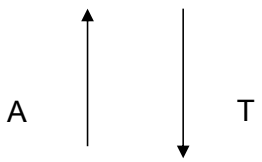
with

$$P = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{18}$$

The relations P and S anticommute:

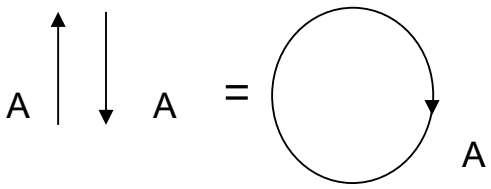
$$PS + SP = 0 \tag{19}$$

We may obtain a better understanding of the relational dynamics by using a double-line representation of the general relation R_{ij}. Each distinct state or base $|B_i\rangle$ is represented by a specific line, with a downward (upward) arrow attached to the initial (final) state. In this sense we picture $R_{14} = |B_1\rangle \langle B_4|$ by



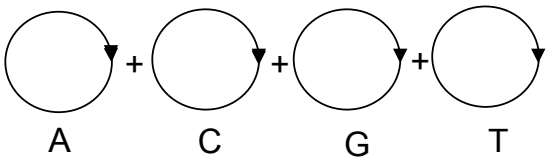
(20)

The relation $R_{11} = |B_1\rangle \langle B_1|$ is pictured by



(21)

The identity, eq. 11, appears as a collection of bubbles



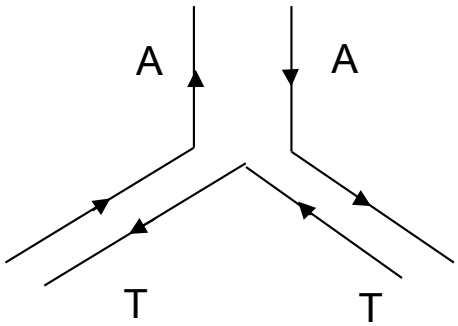
(22)

We consider that the identity expression and its picture represent the DNA vacuum. The DNA vacuum is not empty but it is full of DNA bases, which are continuously annihilated and created. This restless DNA vacuum resembles the QCD vacuum. QCD (Quantum Chromodynamics) describes the strong interactions among quarks, the fundamental entities of hadrons (proton, neutron, pion etc.). The QCD vacuum is a sea of restless quarks and antiquarks.

We may move further and picture the composition of relations, eq. (1). For example, the composition

$$R_{14}R_{41} = R_{11} \tag{23}$$

will look like



(24)

Clearly the above diagram looks like a string diagram, but at the same time it represents the DNA replication process starting from the DNA vacuum. Next to the discrete operation R, we may consider a continuous operation $e^{\alpha R}$, where α is a parameter. A Taylor expansion provides

$$e^{\alpha R} = \mathbf{1}\cos\alpha + R\sin\alpha \tag{25}$$

The operation $e^{\alpha R}$ acts like a rotation. Consider the bonding $|T\rangle\langle A|$. Upon rotation it is transformed to

$$\begin{aligned} e^{\alpha R}\left[|T\rangle\langle A|\right]e^{\alpha R} &= \cos^2\alpha\,|T\rangle\langle A| + \\ &+\sin^2\alpha\,|A\rangle\langle T| + \\ &+\frac{1}{2}\sin 2\alpha\left[|T\rangle\langle T| + |A\rangle\langle A|\right] \end{aligned} \tag{26}$$

For $\alpha = \frac{\pi}{4}$ the final result becomes

$$\frac{1}{2}\left[|T\rangle\langle A| + |A\rangle\langle T| + |T\rangle\langle T| + |A\rangle\langle A|\right] \tag{27}$$

We observe that next to the initial bonding $|T\rangle\langle A|$, we obtain in a formal way the bonding $|A\rangle\langle T|$, plus a collection of "sea" DNA bases. Thus the replication process is a direct manifestation of the Watson-Crick complementarity principle.

3. Gödels numbering for DNA

A single DNA strand is a chain of bases, where two consecutive bases are bound together by a covalent bond. We may consider that each of the four bases stands for a letter, establishing a 4-letter alphabet. A succession of these letters in a DNA strand may represent a word or a phrase in a biological language. From another point of view each base may represent a symbol in an axiomatic system. A succession of these symbols may correspond to a mathematical theorem.

Along this line, we are entitled to be inspired by the work of Kurt Gdel [5]. Gdel made an immense impact upon scientific and philosophical thinking in the 20th century, by establishing the incompleteness theorem. To prove this theorem, Gdel developed a technique, now known as Gdel numbering, which codes formal expressions as natural numbers.

The starting point in Gdels numbering is to assign a positive natural number $\#(s)$ to each of the symbol s, in a fixed but arbitrary way. This is the Gdel number of the symbol. We adopt

$$\begin{aligned} &\#(A) = 1,\ \#(C) = 2, \\ &\#(G) = 3,\ \#(T) = 4 \end{aligned} \tag{28}$$

For a succession of DNA bases, for example the four bases AGCA, we pick up the first four prime numbers and we raise each of them to the corresponding Gdel number. Thus

$$w(AGCA) = 2^1 3^3 5^2 7^1 = 9450 \tag{29}$$

is a unique natural number, the Gdel number w for the specific DNA strand. Inversely, given a Gdel number, we can decode it and find the DNA strand it represents. For example, consider the number $w = 3240$. By factorizing w as a product of the prime numbers 2, 3, 5, 7, 11 we may find out how many 2, 3, 5, are hidden in the number. In our example,

$$w = 2^3 3^4 5^1 \tag{30}$$

and therefore the above w represents the DNA strand GTA.

Gdels numbering allows us to obtain a quantitative measure of the difference among the various DNA strands. Consider two strands represented by the Gdel numbers

$$w_1 = 2^{y_1} 3^{y_2} 5^{y_3} 7^{y_4} ... p_k^{y_k} \tag{31}$$

$$w_2 = 2^{q_1} 3^{q_2} 5^{q_3} 7^{q_4} ... p_k^{q_k} \tag{32}$$

where 2, 3, 5, 7, p_k are the prime numbers and $y_i\,(q_j)$ represents the j-th base in the first (second) strand. The difference D among the two strands is

$$D = \frac{w_1}{w_2} = 2^{(y_1/q_1)} 3^{(y_2/q_2)} 5^{(y_3/q_3)} ... p_k^{(y_k/q_k)} \tag{33}$$

We notice though that the identification $\#(s)$ is arbitrary and therefore D cannot serve as an objective measure. Rather, considering for example w_1 as a reference strand, we define the difference Δ between the reference strand and another strand by

$$\Delta = \prod_j \frac{1}{p_j} \tag{34}$$

where p_j stand for all those prime numbers where the DNA bases differ in the corresponding j places. Imagine the evolution of DNA. We may assume that everything started with an "original" DNA, which upon differentiation provided the subsequent plethora of DNA. We expect that the differences are located at late positions, the corresponding p_j are large and correspondingly Δ is small. On the other hand, if the differences are located at the initial places, then Δ is relatively large.

In a DNA "language" we encounter motifs, that is a precise combination of bases which serves a specific function. Assuming that a motif is a combination of n bases, starting at the position k, its Gdel number $M_{k,n}$, is the product of n prime numbers raised to the appropriate powers, representing the corresponding bases

$$M_{k,n} = p_k^{s_i} p_{k+1}^{s_j} ... p_{k+n}^{l} \tag{25}$$

Then, within a huge DNA collection we can search, through data mining techniques, for the existence of the different $M_{k,n}$.

It should be added, that in order to evaluate similarity or difference in discrete hierarchical biological systems, it has been also proposed to use the ultrametric distance, notably in its padic version [6]. This approach has been useful in classifying the codons.

4. Conclusions and Future Research Directions

There is long standing effort for a "quantum biology". It involves how purely quantum effects, like entanglement and coherence, might be of relevance in biological systems [7 12]. In the present work we do not consider DNA as a

quantum system *per se*. Rather, we suggest that DNAs inherent logic is the relational logic, the same logic that appears governing quantum mechanics and string theory. The emphasis is not in terms of structure, but in terms of information processing. The four DNA bases may be considered as the "letters" of a DNA alphabet and a DNA strand as a "word" or "sentence" in a biological text. The analogy with quantum logic prompted us to use the quantum formalism in order to describe the DNA dynamics. As a by-product of this approach we obtained the DNA vacuum, a collection of annihilated and created DNA bases. An important relation for the DNA is what we defined as the relation of complementarity R. Relation R guarantees the bonding AT and GC and generates the replication process. We have also shown that the relations introduced may be represented by matrices and therefore the whole DNA functioning can be accounted for by matrix mechanics. Within the spirit of matrix mechanics we can search for interactions among matrices a distance apart, giving rise to a sort of "entanglement".

The search of coding regions in a DNA sequence encouraged us to use Gdels numbering in order to codify a DNA strand. This numbering allows to study the existence of a specific DNA "word" within a broader DNA sequence by a simple factorization process. A remarkable feature of languages is Zipfs law. This law dictates that the frequency f of each word in a text and its rank are related according to a power law. It would be most interesting to check if Zipfs law prevails in a biological text [13]. The possible correlation between two DNA strands can be approached also in a novel way. Denoting by x, y the Gdels numbers of the strands, the implicit correlation may be expressed by a function $F(x,y)$, with arguments of the function the numerals x, y.

Acknowledgement

This work was presented at the Belgrade BioInformatics Conference 2016 and the author would like to thank the organizers for the warm hospitality. He would like also to thank Kosmas Kosmidis, Dimitri Evangelinos and Vasilis Niaouris for helping in writing up the paper.

References

1. A. Nicolaidis, "Categorical Foundation of Quantum Mechanics and String Theory",*Int. J. Mod. Phys.* A24: pp. 11751183, 2009.
2. A. Nicolaidis, "Relational Quantum Mechanics, Advances in Quantum Mechanics", Prof. Paul Bracken (Ed.), *InTech*, DOI: 10.5772/54892, 2013.
3. C. S. Peirce, "Description of a notation for the logic of relatives, resulting from an amplification of the conceptions of Booles calculus of logic", *Memoirs of the American Academy of Sciences* 9, pp. 317378, 1870.
4. C. S. Peirce, "On the algebra of logic", *American Journal of Mathematics* 3, pp. 1557, 1880.
5. Gdel, K., 1931, ber formal unentscheidbare Stze der Principia Mathematica und verwandter Systeme I, *Monatshefte fr Mathematik Physik*, 38: pp. 173198. English translation in "Gdel Collected Works I". Publications 19291936, S. Feferman et al. (eds.), Oxford: Oxford University Press, pp. 144195, 1986.

6. B. Dragovich and A. Dragovich, “p-Adic Modelling of the Genome and the Genetic Code”, *Computer Journal* 53, 2010.
7. M. Arndt, T. Juffmann and V. Vedral, “Quantum physics meets biology”, HFSP journal, vol. 3, pp. 386400, 2009.
8. N. Lambert, Y-N Chen, C-M Li, G-Y Chen and F. Nori, “Quantum biology”, *Nature Physics*, vol. 9, Jan. 2013.
9. G. Vattay, S. Kauffman and S. Niiranen, “Quantum biology on the edge of quantum chaos”, *PLOS one*, vol. 9, issue 3, e89017, March 2014.
10. I. Kominis, “The radical-pair mechanism as a paradigm for the emerging science of quantum biology”, *Mod. Phys. Lett.* B, 29, 1530013, 2015.
11. L. Kauffman, “Self-reference, biologic and the structure of reproduction”, arXiv:1512.04325, 2015.
12. D. V. Nanopoulos, “Theory of Brain Function, Quantum Mechanics and Superstrings”, invited talk at the Physics without frontiers Four Seas Conference, Trieste, hep-ph-9505374, 1995.
13. A. Tsonis, P. Kumar, J. B. Elsner, and P. A. Tsonis, “Is DNA a Language?”, *J. Theor. Biol.* 184, pp. 2529, 1997.

Mining PMMoV genotype-pathotype association rules from public databases

Vesna Pajić[1], Bojana Banović[2], Miloš Beljanski[3] and Dragana Dudić[1]

[1] Center for Data Mining and Bioinformatics, Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11080 Zemun, Serbia
{svesna, ddragana}@agrif.bg.ac.rs
[2] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11000 Belgrade, Serbia
bojanabanovic@imgge.bg.ac.rs
[3] Institute for General and Physical Chemistry, University of Belgrade, Studentski Trg 12, 11158 Belgrade, Serbia
mbel@matf.bg.ac.rs

Abstract. . In order to utilize knowledge hidden in public databases, we applied several data mining techniques on PMMoV sequences from NCBI nucleotide database with an aim to characterize this virus at molecular level. The dataset consists of 231 nucleotide sequences collected. We identified three distinct genotype variants (namely TG, GA and GG) based on the nucleotide combinations on significant positions within subgroups of sequences. Those positions were further confirmed using the EM algorithm. The information about pathotype was known for only 40% of studied sequences and distribution of pathotypes was very imbalanced. Nevertheless, using the Apriori-type algorithm two strong rules was mined (confidence 0.96 and 0.93). The analysis showed that hidden knowledge could be disclosed and put to use through data mining approaches like class association analysis and cluster analysis.

Keywords: clustering, class association rules, PMMoV

1. Introduction

With new sequencing technologies, field of genomics is growing fast and so is the amount of the data behind it. Most of that data is publicly available through different data sources in a recent molecular biology databases review [1] there are even 1685 relevant resources in molecular biology reported, where each data source contains a large amount of data. NCBI nucleotide database contains sequences from multiple sources including GenBank with 190,250,235 sequences[4], RefSeq with 92,936,289 sequences[5], and PDB with 117,240 sequences[6]. Although a vast amount of sequence data is available, there is a huge and mostly unrealized potential in analyzing it.

[4] http://www.ncbi.nlm.nih.gov/genbank/statistics/
[5] http://www.ncbi.nlm.nih.gov/refseq/statistics/
[6] http://www.rcsb.org/pdb/statistics/holdings.do

In this research we choose Pepper Mild Mottle Virus (PMMoV) as *in silico* plant virus model to test what kind of information one can extract from publicly available nucleotide sequences by using bioinformatics tools and data mining approach. PMMoV is a tobamovirus responsible for diminishing pepper yields. Until 2005, soil treatment against PMMoV consisted of the application of methyl bromide, ozone depleting chemical. By utilizing publicly available PMMoV's sequence data one could explore life cycle, pathogenicity, virulence potential and plant resistance mechanisms of the virus in order to develop more eco-friendly alternatives for the suppression of the virus in the field. We analyzed nucleotide content and single nucleotide variations of available sequences with several data mining techniques, and compared the results with information on virus pathogenicity found in the literature [2–4]. The overall aim was to detect some of existing relations between nucleotide content and pathotype which could potentially be used for future monitoring of virus and its pathogenicity.

2. Data

At the time of the analysis, 231 PMMoV nucleotide sequences were available in NCBI database at total; 13 of them were complete genomes, 150 corresponded to coat protein, 62 corresponded to 126K replicase small subunit, 6 corresponded to 183K replicase large subunit and 7 corresponded to 30K cell-to-cell movement protein. They constituted dataset D1 and were aligned using Clustal X 2.1[7], with later manual correction in MEGA 6[8]. We used package *seqinr* in R in order to determine profile sequence, in respect of which all other analyses were conducted.

There were 94 sequences (40%) in the dataset D1 for which information on pathotype (one of five pathotypes: P_0, P_1, P_{12}, P_{123} and P_{1234} described in literature [5, 6]) was available either in papers or in NCBI database. For the purpose of mining genotype-pathotype association rules, these sequences, along with the information about genotype (determined in this research) and pathotype were extracted in another dataset, the dataset D2.

Dataset D1 was additionally split into groups and subgroups based on the part of the genome the sequences were covering (Table 1). The whole genome sequences were then divided into subsequences corresponding to the nucleotide positions the subgroups were covering, so each subgroup had got 13 more sequences, obtained from the whole genome sequences. The positions covered overlapped for some subgroups; for example sequences from subgroup 1.1 covered 200 positions (from 612 to 810), and sequences from subgroup 1.2 span across these positions, but also covered additional bases (from 481 to 1248).

[7] http://www.clustal.org/clustal2/
[8] http://www.megasoftware.net/

TABLE 1: Number of sequences and nucleotide positions (np) for each subgroup

Group	Subgroup	Number of sequences	Number of np covered	First np in the genome	Last np in the genome
1	1.1	67	200	612	810
	1.2	50	768	481	1248
2	2	22	190	1622	1811
3	3	18	790	4015	4805
4	4.1	15	779	4909	5682
	4.2	110	476	5685	6157
	4.3	160	209	5841	6047

3. Tools, Methods and Algorithms

For fulfilling data mining tasks we used WEKA 3.6.10[9] algorithm implementations. Bioinformatics analyzes were performed using Bioconductor package in R. After aligning sequences of each subgroup, single nucleotide variations (SNVs) were determined with in-house script. Comprehensive analysis of SNVs revealed several informative nucleotide positions in each subgroup. Based on the combination of nucleotides contained in these positions, sequences could be divided into disjoint sets in a way that each sequence from a set contains the same, unique combination of nucleotides on the positions Information about the sets, determined significant positions and nucleotide combinations on them is shown in Table 2.

Analysis of relationships among these sets, for the sequences spanning in more than one group (explained further in Section 4.2, was used for the detection of three distinct genotype variants, which were additionally confirmed with cluster analysis.

3.1. Clustering

Probabilistic models are often used for modeling biological data but nowadays, when produced sequence data are noisy, incomplete and erroneous, it is difficult to determine parameters of a model. To estimate parameters in a probabilistic models with incomplete data, Expectation Maximization (EM) algorithm [7] is used. In order to confirm determined genotype variants and to disclose other existing similarities between sequences, we performed cluster analysis on D1 dataset using EM algorithm. The optimum number of clusters was estimated via 10-fold cross validation.

3.2. Class Association Rules

The Apriori algorithm [8] is a widely used essential technique for learning association rules from a given dataset with minimum support and minimum confidence parameters specified. We used modification of the original algorithm

[9] http://www.cs.waikato.ac.nz/ml/weka/

TABLE 2: Nucleotide positions that divide sequences into disjoint sets in each group

Group	np in the genome sequence	Set label	Nucleotide combination (short mark)*	Number of sequences
1.1	639; 669	G1.1-1	GT (G)	52
		G1.1-2	AC (A)	15
1.2	565; 566; 708;1125	G1.2-1	TGGA (T)	34
		G1.2-2	GTTG (G)	16
2	1638; 1647	G2-1	TA (T)	12
		G2-2	CG (C)	9
3	4107; 4131; 4392; 4395;	G3-1	GACGTCCA (G)	10
	4516; 4560; 4650; 4698	G3-2	AGTACTTG (A)	8
4.1	4929; 4963; 5085; 5151;	G4.1-1	CGACACGG (C)	10
	5244; 5487; 5557; 5611	G4.1-2	TAGGGTAA (T)	5
	5763; 5819; 5837; 5996;	G4.2-1	CTTACTGATGC (C)	86
4.2	6002; 6011; 6038; 6062;	G4.2-2	TCCTTCTTATT (T)	24
	6100; 6101; 6127			
4.3	5996; 6002; 6038	G4.3-1	ACG (A)	127
		G4.3-2	TTT (T)	35

* Short marks are made from the first nucleotide letter in unique nucleotide combination for that set, and they are used just for a shorter representation in text

which combines association rule technique with classification rule technique [9] to allow the algorithm to focus on association rules useful to determine predefined classes. As input parameters we used the dataset D2, minimum support 0.1 and minimum confidence 0.9.

4. Results

4.1. Determination of Genotypes

We analyzed the relationship among sets G1.1-1, G1.1-2, G1.2-1 and G1.2-2 for 50 sequences in Group 1 covering subgroups 1.1 and 1.2. Three distinct genotype variants were determined based on the nucleotide content on sites 565, 566, 639, 669, 708 and 1125: Genotype variant GA, Genotype variant TG and Genotype variant GG.

Evaluation using the EM algorithm resulted with the three clusters in subgroup 1.1, based on already emphasized positions 639 and 708, corresponding to determined genotype variants. In subgroup 1.2, four clusters were mined, one of them contained only one sequence with Genotype variant GG. Distinction of three remaining clusters was based upon 565 np and 552 np. The separation at the earlier stressed position 565 extracts Genotype variant TG. For the clusters formed upon the new obtained position 552 we can state that all sequences from one cluster were having Genotype variant GG and all sequences having Genotype variant GA was in the other cluster, which also contained sequences having Genotype variant GG.

Cluster analysis of the sequences from Groups 2, 3 and subgroup 4.1 did not reveal any new similarities among sequences, but in the subgroup 4.2 it revealed three clusters which corresponded to defined genotype variants. The three clusters obtained using the EM algorithm segregated 6002 np, based on which Genotype variant GA is matched, and newly observed 5975 np which can be used to distinguish Genotype variant GG from Genotype variant TG.

Assuming that revealed information about Group 1 can be transferred to whole genomes (and therefore to isolates) we classified the whole genome sequences based on the determined genotype variants (Table 3).

TABLE 3: Distribution of whole genome sequences into disjoint sets determined by Groups 1.1 – 4.3 analysis and the determined genotype variants (for simplified representation, short marks are used instead of set's labels)

Sequence	1.1	1.2	2	3	4.1	4.2	4.3	Genotype
Spain-1989-P12|NC_003630.1	G	T	T	G	C	C	A	TG
Spain-1989-P12|M81413.1	G	T	T	G	C	C	A	TG
Japan-2005-P0|AB113117.1	G	T	T	G	C	C	A	TG
Japan-2005-P0|AB113116.1	G	T	T	G	C	C	A	TG
Japan-2002-P0|AB069853.1	G	T	T	G	C	C	A	TG
Japan-1997-P1234|AB000709.2	G	T	T	G	C	C	A	TG
China-2006|AY859497	G	T	T	G	C	C	A	TG
Brasil-2010|AB550911.1	G	T	T	G	C	C	A	TG
India-2014-P12|KJ631123.1	G	T	T	G	C	C	A	TG
Japan-2003-P1234|AB276030.1	G	G	G	A	T	C	A	GG
SouthKorea-2005|AB126003.1	G	G	G	A	T	C	A	GG
Japan-2007-P12|AB254821.1	G	G	G	A	T	C	A	GG
Spain-2002-P123|AJ308228	A	G	G	A	T	T	T	GA

4.2. Genotype pathotype associations

The information about pathotype was not available for all sequences from dataset D1 so we could look for genotype pathotype associations only for 94 sequences (dataset D2). The dataset D2 consisted of a sequence name field, genotype variant of the sequence (TG, GG or GA) and pathotype information (denoted as 0, 1, 2, 3 or 4 for pathotypes P_0, P_1, P_{12}, P_{123} and P_{1234} respectively). Applying class association analysis in Weka, we found two strong rules (with high confidence):

$$\text{Genotype variant=TG 45} \Longrightarrow \text{Pathotype=2 43 conf:(0.96)}$$
$$\text{Genotype variant=GA 29} \Longrightarrow \text{Pathotype=3 27 conf:(0.93)}$$

The rules clearly indicate that almost all (43 out of 45) sequences that have Genotype variant TG also have pathotype P_{12}, and that 27 out of 29 sequences having Genotype variant GA also have pathotype P_{123}.

For the sequences having Genotype variant GG, 5921 np (found with a classification method) can be discriminative for pathotype prediction: if sequence has T or C it is of pathotype P12, while if sequence has A it is of pathotype P_{123}.

5. Conclusion

The clustering and class association analysis of 231 PMMoV sequences available at NCBI showed some regularities, not reported previously, which potentially can be used for molecular monitoring of virus genotype-pathotype association. Moreover, our research demonstrated how data mining techniques can be used on publicly available data in order to utilize them more and to extract hidden knowledge that resides in databases.

References

1. Rigden D. J., Fernndez-Surez X. M., Galperin M. Y.: The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. Nucleic Acids Research, 44, D1D6. (2016)
2. Gilardi P, Wicke B, Castillo S, de la Cruz A, Serra MT, Garca Luque I. Resistance in Capsicum spp. against the tobamoviruses. In: Pandalai SG, ed. Recent research developments in virology, Vol. 1. India: Transworld Research Network, 547-558. (1999)
3. Genda Y, Kanda A, Hamada H, Sato K, Ohnishi J, Tsuda S. Two amino acid substitutions in the coat protein of Pepper mild mottle virus are responsible for overcoming the L4 gene mediated resistance in Capsicum spp. Phytopathology 97, 787793. (2007)
4. Antignus, O. Lachman, M. Pearlsman, L. Maslenin, A. Rosner. A new pathotype of Pepper mild mottle virus (PMMoV) overcomes the L4 resistance genotype of pepper cultivars. Plant Dis. 92, 10331037. (2008)
5. Boukema I. W. Resistance to TMV in Capsicum chacoense Hunz. is governed by allele of the L-locus. Capsicum Newsl. 3, 4748 (1984)
6. Sawada H., Takeuchi S., Hamada H., Kiba A., Matsumoto M., Hikichi Y.: A new tobamovirus-resistance gene L1a, of sweet pepper (Capsicum annuum L.). J. Jpn. Soc. Hortic. Sci. 73, 552-557 (2004)
7. Dempster A. P., Laird N. M., and Rubin D. B., Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society B 39: 138. (1977)11.
8. Atluri, G., Gupta, R., Fang, G., Pandey, G., Steinbach, M., Kumar, V., Association analysis techniques for bioinformatics problems, Bioinformatics and Computational Biology, 1-13. (2009)
9. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules, Proceedings of the 20th International Conference on Very Large Databases, 487499. (1994)

Intermittency-driven complexity in the brain: towards a general-purpose event detection algorithm

Paolo Paradisi[1,2], Marco Righi[1], Umberto Barcaro[1], Ovidio Salvetti[1], Alessandra Virgillito[3], Maria Chiara Carboncini[3], and Laura Sebastiani[3]

[1] Institute of Information Science and Technologies (ISTI-CNR), Via G. Moruzzi 1, I-56124 Pisa, Italy
[2] Basque Center for Applied Mathematics(BCAM), Alameda de Mazarredo 14, E-48009 Bilbao, Basque Country, Spain
[3] Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Via Savi 10, 56126 Pisa, Italy
paolo.paradisi@cnr.it

Abstract. In this work we first discuss a well-known theoretical framework for the analysis and modeling of self-organized structures in complex systems. These self-organized states are metastable and rapid transition events mark the passages between self-organization and background or between two different self-organized states. Thus, our approach focuses on characterizing and modeling the complex system as a intermittent point process describing the sequence of transition events.
Complexity is usually associated with the emergence of a renewal point process with power-law distributed inter-event times, hence the term *fractal intermittency*. This point process drives the self-organizing behavior of the complex system, a condition denoted here as *intermittency-driven complexity*.
In order to find the underlying intermittent birth-death process of self-organization, we introduce and discuss a preliminary version of an algorithm for the detection of transition events in human electroencephalograms. As the sequence of transition events is known, the complexity of the intermittent point process can be investigated by applying an algorithm for the scaling analysis of diffusion processes driven by the intermittent process itself. The method is briefly illustrated by discussing some preliminary analyses carried out on real electroencephalograms.

Keywords: signal processing, complexity, fractal intermittency, brain, electroencephalogram (EEG), disorders of consciousness

1. Introduction

The brain is composed of many elementary units, neurons and astrocytes, with an extremely complicated topology of the links among units (axon, dendrites, metabolic network)[4] [1]. The links are characterized by strong nonlinear interactions among neurons (e.g., the chemically activated electrical discharges

[4] Astrocytes are responsible for the regulation of the neural metabolism and, thus, for the energy delivery and storage that neurons need for their electrical activity. The

through the ionic channels) with very complicated feedback mechanisms. The overall picture is that of a complex network with a huge number of nodes (neurons and astrocytes) and links with a very complicated topology. The nonlinear dynamics of single neurons (i.e., the threshold mechanism for the electrical discharges generating spikes and bursts) are highly enhanced by the complex network topology, but at the same time some kind of ordering, or self-organizing, principle triggers the emergence of global cooperativity.
It is then not surprising that brain dynamics display a very rich landscape of different behaviors and a very efficient plastic behavior, characterized by a rapid and efficient capability of response to rapid changes in the external environment. Thus, a great interest is nowadays focused on a better understanding of the brain information processing, with the challenging goal of describing brain complexity by means of a relatively low number of parameters. This is not only a very fascinating problem and a very hot topic in brain research, but it also has important potential applications in several fields (e.g., clinical applications, new diagnostic indices).
In this general framework, the *complexity* approach [2, 3] is focused on the study of emerging self-organized structures in multi-component systems and complex networks. This general approach is nowadays gaining momentum in the field of biomedical signal processing. In order to extract useful information from large clinical datasets, storing many different physiological data and signals, algorithms for the reduction of data complexity are needed to derive reliable diagnostic indices. Then, a great interest is focused in defining, developing and testing statistical indices that can enclose the minimal information required to interpret the basic features of physiological signals. The availability of large datasets storing many different physiological data and signals is asking for reliable procedures of complexity reduction in large datasets. This is needed to extract useful information from the data themselves, which is of much relevance in clinical activity, such as in the diagnosis and treatment of disorders of consciousness (DOC) [4]. However, such indices are useful if they are able to describe the key features of the signals and if these features can be exploited by physicians in their clinical activity, e.g., in the evaluation of a medical condition or disease (diagnosis); in foretelling the course of a disease (prognosis); in the consequent choice of the proper therapy (decision making).

In this work we introduce and discuss an approach to the processing of ElectroEncephaloGrams (EEGs) that is based on the observation that, in many complex systems, such as the human physiology, the nonlinear dynamics of the network trigger intermittent events, each one associated with the emergence or decay of self-organized structures and/or with the transition among different self-organized states [5–12]. Our approach to the modeling of such complex systems and, consequently, to the associated algorithms for data analysis and signal processing is based on the idea that intermittent events drive the complex (self-organizing) behavior of multi-component systems.

role of the substrate network of astrocytes is nowadays recognized to play a crucial role in brain information processing, as it has been recently found that the metabolic component of the brain is characterized by an intense cooperativity between astrocytes and neurons [1].

The paper is organized as follows. In Section 2 we present and discuss the concept of intermittency-driven complexity. After discussing the concept of fractal intermittency (FI), we give a brief review about the emergence of FI in the human brain. In Section 3 we sketch our proposal of a preliminary version of a general-purpose event detection algorithm. Finally, in Section 4, in order to illustrate the event detection algorithm, we briefly discuss an application to real EEG data by showing a few preliminary results.

2. Intermittency-driven complexity

Following the paradigm of *emerging properties*, the complexity approach focuses on the analysis and modeling of self-organized large-scale structures or states emerging from the cooperative dynamics of complex networks. The main idea is that self-organized structures are the essential actors in the global dynamics of complex systems and play a crucial role in many aspects, such as the transport properties and the way the system respond to external stimuli. As a consequence, also the statistical indicators extracted from complex data analysis usually refer to some global property associated with the dynamical evolution of large-scale, global, self-organized states.

2.1. Complexity and fractal intermittency

As far as we know, a general agreement on the definition of complexity does not yet exist. However, we refer here to a class of complex systems displaying the following properties:

(1) a complex system is multi-component with a large number of degrees of freedom, i.e., many functional units or nodes. As said above, these units interact with each other and their dynamics are strongly nonlinear;

(2) non-linearity and multi-component is not enough to define complexity: the dynamics are cooperative and trigger the emergence of self-organized structures, being self-organization not related to the presence of a internal master or to an external ordering force;

(3) self-organized states display long-range space-time correlations (slow power-law decay) and self-similarity (mono- or multi-scaling);

(4) self-organized states are metastable, with relatively long life-times τ and fast transition events between two successive states.

In summary, the cooperative dynamics determine an alternation of strongly correlated self-organized structures and a background characterized by short-term correlations, or an alternation among different self-organized states. The passages are marked by fast transitions that can be considered quasi-instantaneous events. The n-th event occurs at a random time t_n. The sequence of transition events is an emergent property described as a a intermittent birth-death point process of self-organization: $\{t_n\}_{n=0}^{N}$; $t_0 = 0$. Then, in the above list the feature (4) is a crucial one as it allows for a description of complexity in terms of intermittent events.

Due to the fast memory drop occurring during the fast transitions, each self-organized state is often independent from each other, as such as the crucial transition events and the inter-event times, also named Waiting Times (WTs): $\tau_n = t_n - t_{n-1}; n = 1, 2, ...$ This is denoted as *renewal condition*. In this case, the sequence of crucial events is described by a renewal point process. A complex (cooperative) system is characterized by metastable self-organized states whose life-times τ_n are statistically distributed according to a inverse power-law function. This condition, i.e., the triggering of fast transition events that are renewal and with inverse power-law WT distribution, is denoted as *fractal intermittency* (FI) [9, 13–16]. The term *intermittency-driven complexity* (IDC) is here used to indicate both the associated complex behavior and the class of complex systems displaying FI. In this case the birth-death point process of self-organization is given by a non-Poisson process (renewal or not). The departure from the Poisson reference condition is a signature of complexity. In fact, a Poisson (renewal) point process is typically associated with the lack of cooperation and self-organizing behavior. A Poisson process does not generate neither long-range correlation or memory nor fractal intermittency, being the WT distribution given by an exponential decay [13, 15, 16]. Despite the presence of renewal events, the autocorrelation function of the intermittent signal can be long-range, i.e., with a slow power-law decay in the tail: $C(t) \sim 1/t^{\beta}$ [5].
The correlation exponent β is an important example of *emergent property* that can be used as a synthetic indicator of the cooperative dynamics in the complex system. For a complex system in the IDC class, β is related to the power index μ of the inverse power-law in the WT distribution:

$$\psi(\tau) \sim \frac{1}{\tau^{\mu}} \ . \tag{1}$$

Analogously to β, also $\psi(\tau)$ and μ are *emerging properties* and, thus, a signature of complex behavior. The parameter μ, denoted as *complexity index*, is an example of a statistical index that can quantify the presence of IDC in a system, thus evaluating the ability of the complex system to trigger self-organization. Other indices, depending on μ, can also be used as IDC indices [9, 13–16]. In particular, complexity is identified with a condition of very slow decay in $\psi(\tau)$, corresponding to the range $1 < \mu < 3$ (see Refs. [13, 16] for details). In Fig. 1 we sketch a synthetic scheme qualitatively explaining the connection between self-organization, cooperation and non-Poisson renewal processes. Poisson renewal processes always emerge in the case of independent systems, whatever the microdynamics of the single nodes. As a consequence, a departure from the Poisson statistics reveals some kind of cooperation among the nodes of the network. Further, the emergence a fractal renewal process, i.e., a renewal process with inverse power-law WT distribution (fractal intermittency), means that cooperation is complex, i.e., associated with complex self-organized structures.

2.2. Fractal intermittency in the brain: a brief survey of results

Metastability is a basic feature of the information processing in the brain neural network. Fingelkurts and Fingelkurts recognized that rapid changes in EEG

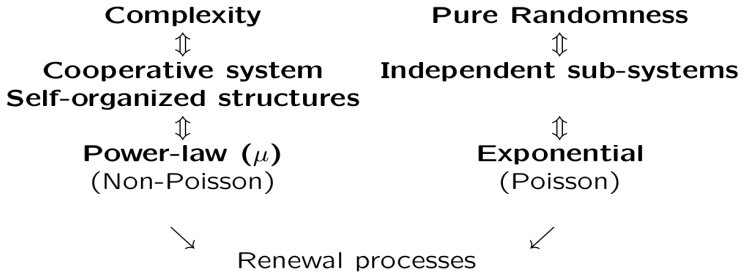


FIG. 1: Comparison of Poisson (non-complex) and non-Poisson (complex) renewal processes.

records, called Rapid Transition Processes (RTPs), mark passages between two quasi-stationary periods, each one corresponding to different neural metastable assemblies, and are the signature of brain self-organization [17, 18]. Neural assemblies are associated with transient information flow among different neurons with the goal of developing a specific brain function and/or the response to external stimuli (e.g., Event Related Potentials). RTPs and neural assemblies are then a prototype of crucial events and meta-stable self-organized states, respectively.

The algorithm for the automatic detection of RTP events in EEG data was developed in Ref. [18] and exploited by the authors of Refs. [5, 7, 6, 8–11] to characterize the complexity of the intermittent events. By exploiting a scaling detection method, the Event-Driven Diffusion Scaling (EDDiS) algorithm (see [13] and references therein), these authors found that brain dynamics display fractal intermittency. In particular, it was shown that the fractal intermittency approach is able to reveal the integrated (Rapid Eye Movement, REM) and segregated (Non-REM) stages during sleep, thus in agreement with the consciousness state of the subjects [9–11]. This important result proves that the IDC concept and the associated IDC measures could be good candidates for the characterization of DOCs.

In the intermittency-based analysis here proposed, a key aspect is the definition of events, which needs to be further studied in order to extend the above analysis to different experimental and clinical conditions.

3. Intermittency-based processing of complex physiological signals

The results obtained by applying the algorithms cited above, the RTP event detection algorithm [18] and the EDDiS algorithm [5, 16, 15, 13], are very promising in the perspective of potential applications in the clinical activity of neurological disorders. However, RTP events are defined only for some experimental conditions.

In this work we investigate the key aspect of the event definition. We propose an algorithm involving a more general definition of event and being able to detect

and discriminate events with different neuro-physiological origins. The proposed method essentially extends the technique introduced and applied in Refs. [19–21], which allows to extract different kind of events marking the sudden increase of activity in given frequency bands.

We assume that the signals were already pre-processed for the artifact cleaning. Then, the event detection algorithm works as follows:

(1) Splitting of the single EEG channel into different frequency bands.
The following band ranges are usually considered: (a) δ band ($0.5-4$ Hz); (b) θ band ($4-8$ Hz); (c) α band ($8-12$ Hz); (d) σ band ($12-16$ Hz); (e) β band ($16-35$ Hz); (f) γ band ($35-64$ Hz).

(2) For each frequency band, the component amplitude (the absolute value) is considered. Then, two moving-window time averages are computed at different time scales, a short and a long one, being this last one used to evaluate the signal envelope[5].

(3) Calculation of non-dimensional descriptors $A_k(t)$ for each frequency band $k=\delta,\theta,...$: (short-time average - long-time average)/long-time average. Then, the global average $\overline{A}_k$ (or some local average) of $A_k(t)$ is computed and subtracted to $A_k(t)$ itself: $S_k(t)=A_k(t)-\overline{A}_k$.

(4) Identification of high- and low-activity epochs and of transition events between epochs. This is done for each frequency band by using a thresholding technique, whose details and parameters can also be changed depending on the specific events that must be detected. The most simple method, which be applied in the next section, is given by the zero crossings of the $S_k(t)$[6].

(5) Storing in a database (spatio-temporal event maps).
Extraction of specific kinds of events from the event maps.

(6) Feature extraction from the event sequences and maps, such as: number of events per time unit for each band and/or EEG trace; covariance matrix based on events; estimation of complexity index, both for single EEG channels and global events (temporal coincidences among different EEG channels).

Despite its apparent simplicity, this algorithm is very flexible and powerful. Being based on the classical Fourier approach and on splitting the EEG signal into standard frequency bands, this approach allows for a more clear link between the event detection algorithm and its neuro-physiological interpretation. In this sense, a particular kind of brain events should be recognized to be a neural correlate of some increased or decreased neurophysiological activity.

[5] In the original applications of the method [19–21] typical chosen values of the averaging times were 2 and 64 sec., as these values were the most suitable to detect macroscopic epochs of high intensity in the given frequency band.

[6] The particular definition of event remain the most subtle point of the IDC approach. As an example, if we are interested in characterizing the epochs with substantially increased activity in a given band, and the associated transition events from/to these same epochs, usually two thresholds are used, a low and a high one. This is the standard approach used to automatically detect the waveforms that could be investigated also by visual inspection [19–21].

4. An application to EEG data: preliminary results

In this final section we briefly illustrate an application of the event detection algorithm to real EEG records. The EEG data were collected during a study performed in the Brain Injury Unit, Department of Neuroscience, Cisanello Hospital, Pisa, Italy [22]. A few unconscious patients were treated with a drug (Zolpidem)[7], with the working hypothesis that Zolpidem might increase the brain activity in a rapid way (i.e., within 30 minutes). The general objective of this pilot study was to stimulate a rapid recovery of the patient's consciousness. Clear clinical evidence was not obtained and it would be desiderable to get some kind of indications that the single treatment had some kind of effect on the brain electrical activity.
Here we show a preliminary analysis on one patient, whose EEG was recorded according to the international $10-20$ configuration system. The EEG was recorded

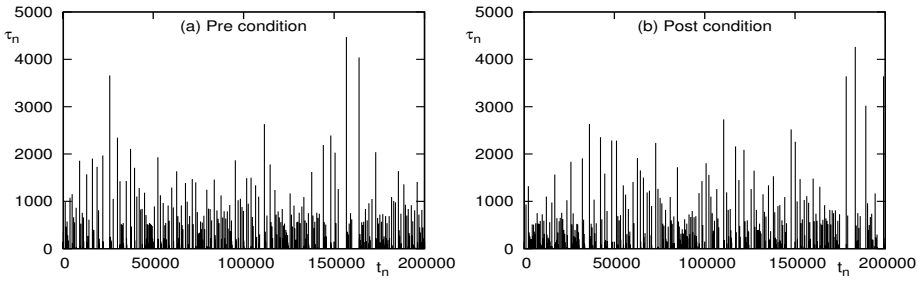


FIG. 2: Waiting Times τ_n vs. the occurrence times t_n of transition events (zero crossings) for the α band of the O_2 electrode. *Pre* condition in Panel (a) refers to the EEG baseline before the Zolpidem treatment, while *Post* condition in Panel (b) refers to the EEG measured 30 minutes after the Zolpidem treatment.

before (baseline) and 30 minutes after the Zolpidem treatment. The sequence of zero crossing events and the associated WTs are derived for each frequency band. In Fig. 2 we report the sequence of WTs τ_n extracted from the α band of the O_2 electrode. WTs are plotted as a function of the event occurrence times t_n. *Pre* and *Post* conditions are reported in Panels (a) and (b), respectively. In Fig. 3 we show the histograms of WTs extracted from two different bands, α (Panel (a)) and σ (Panel (b)), of the same electrode (O_2). For each frequency band the *Pre* and *Post* conditions are reported and compared.
As a general observation, we can say that no qualitative and/or quantitative difference between the *Pre* and *Post* conditions is clearly stated and the WT distributions appear to be almost identical for the two conditions. This situation is also seen in the other electrodes and frequency bands. However, as said above, IDC is investigated by estimating the anomalous scaling behavior of diffusion processes that are built in a proper way using the transition events extracted

[7] The hospital ethical committee approved the study and Informed written consent was obtained from the guardians or relatives of the patients.

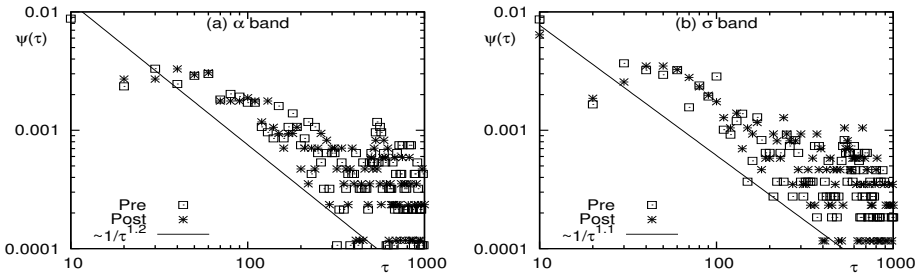


FIG. 3: Histograms $\psi(\tau)$ of the WTs extracted from the electrode O_2 for the bands α (Panel (a)) and σ (Panel (b)). The bin size is 10. The *Pre* and *Post* conditions are compared in each panel. The inverse power-law functions are reported as a guide-to-the-eye.

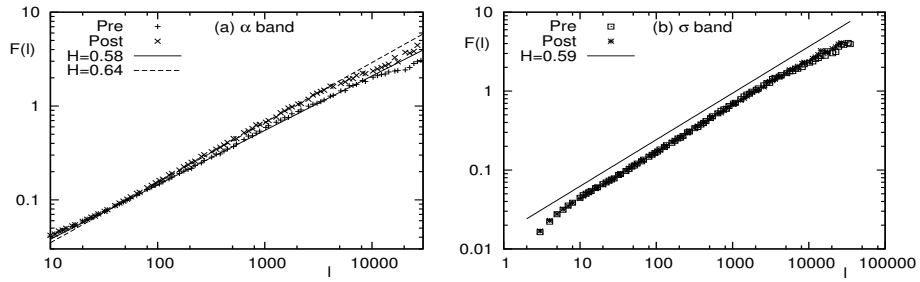


FIG. 4: DFA function $F(l)$ as a function of the time lag l for the O_2 electrode. Panel (a) α band; Panel (b) σ band. The *Pre* and *Post* conditions are compared in each panel. The estimated index μ is also reported.

from the signals. The EDDiS algorithm is applied to the WT sequences in order to obtain one or more (event-driven) diffusion scaling indices and/or the complexity index μ (see Ref. [13] for a review of the EDDiS algorithm). Here we limit ourselves to derive a single diffusion process, which is defined by allowing a random walker to make a unitary jump ahead $(+1)$ in correspondence of each transition event. This walking rule is known as Asymmetric Jump (AJ) and it has been proven to be an efficient and reliable method for the scaling evaluation (see Ref. [13] and references therein, in particular Ref. [23]). Applying the AJ rule we get a diffusing variable $X(t)$. The IDC is estimated through the second moment scaling H, corresponding to the Hurst exponent, which is defined by the following power-law behavior:

$$F(l) \sim l^H; F^2(l) = \langle \left(X^(l) - \overline{X}(l) \right)^2 \rangle \, . \tag{2}$$

Here l is the length of the time window that is moved along the time series in order to carry out the time average. In fact, the statistical analysis of time signals can be only carried out by means of time averages. For any l, the signal is divided into time windows of length l. Each segment is a pseudo-trajectory of a statistical ensemble of path of total duration l. The second moment $F^2(l)$

of properly detrended fluctuations is computed by averaging over this statistical ensemble.
In our specific application, the second moment scaling H is computed by using the Detrended Fluctuation Analysis (DFA) [24], which is a scaling detection algorighm based on a proper evaluation of the trend $\overline{X}(l)$. The value $H = 0.5$, named normal diffusion scaling, indicates the absence of long-range correlations and of network connectivity. Then, the neurons whose electrical activity contribute to the electrode signal (O_2 in this case) do not cooperate each other. Values around $H = 0.5$ indicate low levels of cooperation, so the departure from the normal diffusion scaling is a measure of network cooperativity. When applied to single EEG electrodes this feature can be exploited as a signature of the functional connectivity, i.e., cooperative behavior, of the particular brain region affecting the electrode potential.
In Fig. 4 we show the results of the DFA applied to the diffusing random walk $X(t)$ computed applying the AJ rule to the WTs whose distributions are given in Fig. 3. The diffusion scaling of the σ band (Panel (b)) does not change significantly before and after the Zolpidem treatment. On the contrary, the α band (Panel (a)) shows a small, but net difference between the two conditions. In particular, the IDC index H changes from $H = 0.58$ to $H = 0.64$, which correspond to a small increase in the complexity of the brain dynamics, at least in the region corresponding to O_2. Interestingly, the values of H here estimated for a DOC patient are, as expected, much smaller that the typical values found in conscious healthy subjects, that is, $H \sim 0.75 - 0.95$ [5, 9–11]. Further, it is worth noting that the difference *Pre-Post* cannot be appreciated from the comparison of WT distributions, as it is clearly seen in Fig. 3.
This preliminary analysis and discussion is far from being conclusive and needs further investigations. In particular, a global analysis of the brain network will be carried out. However, in previous papers it has been shown that the IDC indices can characterize the kind of brain connectivity determining the emergence of consciousness [9–11]. Thus, we are convinced that the use of event-driven diffusion scaling analysis (EDDiS) for the investigation of the brain IDC can have potential applications in the field of neurological diseases.

References

1. M. Bélanger, I.Allaman and P.J. Magistretti: Brain Energy Metabolism: Focus on Astrocyte-Neuron Metabolic Cooperation, Cell Metabolism **14**, 724-738 (2011).
2. R. V. Solé, J. Bascompte: Self-organization in Complex Ecosystems. Princeton University Press, Princeton (2006).
3. D. Sornette, Critical phenomena in natural sciences. Springer-Verlag, Berlin (2006).
4. A.A. Fingelkurts, A.A. Fingelkurts, S. Bagnato, C. Boccagni and G. Galardi: Do we need a theory-based assessment of consciousness in the field of disorders of consciousness ?, Front. Hum. Neurosci. **8**(402), doi: 10.3389/fnhum.2014.00402 (2014).
5. P. Allegrini, D. Menicucci, R. Bedini, L. Fronzoni, A. Gemignani, P. Grigolini, B.J. West, P. Paradisi, Spontaneous brain activity as a source of ideal 1/f noise, Phys. Rev. E **80** (2009), 061914.
6. P. Allegrini, D. Menicucci, R. Bedini, A. Gemignani, P. Paradisi, Complex intermittency blurred by noise: Theory and application to neural dynamics. Phys. Rev. E 82 (2010) 015103.

7. P. Allegrini, P. Paradisi, D. Menicucci, A. Gemignani, Fractal complexity in spontaneous EEG metastable state transitions: new vistas on integrated neural activity. Frontiers in Physiology 1, 128 (2010).
8. P. Allegrini, P. Paradisi, D. Menicucci, R. Bedini, A. Gemignani, L. Fronzoni, Noisy cooperative intermittent processes: From blinking quantum dots to human consciousness. J. Phys.: Conf. Series 306 (2011) 012027.
9. P. Paradisi, P. Allegrini, A. Gemignani, M. Laurino, D. Menicucci, A. Piarulli, Scaling and intermittency of brain events as a manifestation of consciousness, AIP Conf. Proc. 1510 (2013), 151-161.
10. P. Allegrini, P. Paradisi, D. Menicucci, M. Laurino, R. Bedini, A. Piarulli, A. Gemignani, Sleep unconsciousness and breakdown of serial critical intermittency: New vistas on the global workspace. Chaos, Solitons and Fractals 55 (2013) 32-43.
11. P. Allegrini, P. Paradisi, D. Menicucci, M. Laurino, A. Piarulli, A. Gemignani, Self-organized dynamical complexity in human wakefulness and sleep: Different critical brain-activity feedback for conscious and unconscious states, Phys. Rev. E **92**, (2015) 032808.
12. P. Grigolini and D.R. Chialvo (Eds.): Emergent Critical Brain Dynamics (Special Issue), Chaos Sol. Fract. **55**, 1-120, Elsevier, Amsterdam (2013).
13. P. Paradisi, P. Allegrini, Scaling law of diffusivity generated by a noisy telegraph signal with fractal intermittency, Chaos, Solitons and Fractals **81** (2015), 451462.
14. P. Paradisi, G. Kaniadakis, A.M. Scarfone, The emergence of self-organization in complex systemsPreface, Chaos, Solitons and Fractals **81** (2015) 407411.
15. P. Paradisi, R. Cesari, A. Donateo, D. Contini, P. Allegrini, Scaling laws of diffusion and time intermittency generated by coherent structures in atmospheric turbulence. Nonlinear Processes in Geophysics 19 (2012) 113-126; P. Paradisi et al., Corrigendum, Nonlinear Processes in Geophysics 19 (2012) 685.
16. P. Paradisi, R. Cesari, A. Donateo, D. Contini, P. Allegrini, Diffusion scaling in event-driven random walks: an application to turbulence. Rep. Math. Phys. 70 (2012) 205-220.
17. A. A. Fingelkurts, A. A. Fingelkurts, Brain-Mind Operational Architectonics Imaging: Technical and Methodological Aspects. Open Neuroimag. J. 2 (2008) 73-93.
18. A.Y. Kaplan, A.A. Fingelkurts, A.A. Fingelkurts, B.S. Borisov, B.S. Darkhovsky, Nonstationary nature of the brain activity as revealed by EEG/EMG: methodological, practical and conceptual challenges. Signal Process. 85 (2005) 2190-2212.
19. C. Navona, U. Barcaro, E. Bonanni, F. Di Martino, M. Maestri, L. Murri, An automatic method for the recognition and classification of the A-phases of the cyclic alternating pattern, Clin. Neurophysio. **113** (2002), 1826-1833.
20. U. Barcaro, E. Bonanni, M. Maestri, L. Murri, L. Parrino, M.G. Terzano, A general automatic method for the analysis of NREM sleep microstructure, Sleep Med. **5** (2004), 567-576.
21. M. Magrini, A. Virgillito, U. Barcaro, L. Bonfiglio, G. Pieri, O. Salvetti, M.C. Carboncini, An automatic method for the study of REM sleep microstructure, Int. Workshop on Computational Intelligence for Multimedia Understanding (IWCIM 2015), Prague, 29-30 October 2015, DOI: 10.1109/IWCIM.2015.7347066 [IEEE Xplore Digital Library]
22. G. Valenza, M.C. Carboncini, A. Virgillito, I. Creatini, L. Bonfiglio, B. Rossi, A. Lanatá and E.P. Scilingo, EEG complexity drug-induced changes in Disorders of Consciousness: a preliminary report, Proceedings of the 33rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC2011), 3724-3727 (2011) [Website: embc2011.embs.org/].
23. P. Grigolini, L. Palatella, G. Raffaelli, Asymmetric anomalous diffusion: an efficient way to detect memory in time series. Fractals 9 (2001) 439-449.

24. C. -K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, A. L. Goldberger, Mosaic organization of DNA nucleotides. Physical Review E 49 (1994) 1685.

A Mathematical description of the Genetic Code: Symmetry and Minimum Principle

A. Sciarrino[1] and P.Sorba[2]

[1] I.N.F.N., Sezione di Napoli
Complesso Universitario di Monte S. Angelo
Via Cinthia, I-80126 Napoli, Italy
nino.sciarrino@gmail.com
[2] LAPTH,Laboratoire d'Annecy-le-Vieux de Physique Théorique CNRS
Université de Savoie
Chemin de Bellevue, BP 110,
F-74941 Annecy-le-Vieux, France
paul.sorba@lapth.cnrs.fr

Abstract. The present paper is a review of the "Crystal Basis Model" of the genetic code proposed a few years ago. The elaboration and verification of sum rules for codon usage probabilities, predictions for some physical-chemical properties of amino-acids as well as some consequences of the application of a "minimization principle" in the mRNA editing and in the codon bias are reviewed.

Keywords: crystal basis model, codon usage frequency, physical-chemical properties of amino acids, codon-anticodon interaction, codon bias

1. Introduction

In the mathematical framework we have proposed [1], the codons appear as composite states of nucleotides. More precisely, the codons are obtained as tensor products of nucleotides, the four nucleotides being assigned to the fundamental representation of the quantum group $\mathscr{U}_q(sl(2)\oplus sl(2))$ in the limit of the deformation parameter $q \to 0$. The use of a quantum group in the limit $q \to 0$ is essential to take into account the nucleotide ordering (see Table 1).

TABLE 1: The eukariotic code

codon	a.a.	codon	a.a	codon	amino acid	codon	a.a.
CCC	Pro P	UCC	Ser S	GCC	Ala A	ACC	Thr T
CCU	Pro P	UCU	Ser S	GCU	Ala A	ACU	Thr T
CCG	Pro P	UCG	Ser S	GCG	Ala A	ACG	Thr T
CCA	Pro P	UCA	Ser S	GCA	Ala A	ACA	Thr T
CUC	Leu L	UUC	Phe F	GUC	Val V	AUC	Ile I
CUU	Leu L	UUU	Phe F	GUU	Val V	AUU	Ile I
CUG	Leu L	UUG	Leu L	GUG	Val V	AUG	Met M
CUA	Leu L	UUA	Leu L	GUA	Val V	AUA	Ile I
CGC	Arg R	UGC	Cys C	GGC	Gly G	AGC	Ser S
CGU	Arg R	UGU	Cys C	GGU	Gly G	AGU	Ser S
CGG	Arg R	UGG	Trp W	GGG	Gly G	AGG	Arg R
CGA	Arg R	UGA	Stop	GGA	Gly G	AGA	Arg R
CAC	His H	UAC	Tyr Y	GAC	Asp D	AAC	Asn N
CAU	His H	UAU	Tyr Y	GAU	Asp D	AAU	Asn N
CAG	Gln Q	UAG	Stop	GAG	Glu E	AAG	Lys K
CAA	Gln Q	UAA	Stop	GAA	Glu E	AAA	Lys K

We have distinguished two parts in this review.

The first one starts with a rapid recalling of the main aspects of our model that we called “Crystal Basis Model”. It is followed by two examples of applications. The first one concerns the setting of sum rules for codon usage probabilities [2]: it is deduced that the sum of usage probabilities of codons with C and A in the third position for the quartets and/or sextets is independent of the biological species for vertebrates. The second application deals with the physical-chemical properties of amino-acids for which a set of relations have been derived and compared with the experimental data [3]. A prediction for the not yet measured thermo-dynamical parameters of three amino-acids is also proposed.

The “minimum” principles, in their different formulations, have played and play a very relevant role in any mathematically formulated scientific theory. The key point of a “minimum” principle is to state that an event happens along the path that minimizes a suitable function.

Keeping at hand the minimisation of our codon-anticodon interaction potential introduced in [4] and [5] for an analysis of the genetic code evolution, codon bias are discussed, providing inequalities between codon usage probabilities for quartets of codons [6]. Performing this study separately for the Early and for the Eukariotic genetic code, we observe a consistency with the obtained results as well as good agreement with the available data. Last but not least, an analysis of the coherent change of sign, in the evolution from the Early to the Eukaryotic code, of the two parameters regulating our interaction potential is performed.

As this paper is essentially a review of the Crystal Basis Model, we have limited the references and only provided those directly connected to our approach. The interested reader can find in each quoted paper the relative biography.

2. PART 1: Crystal basis model and application

2.1. A group theoretical model of the genetic code

We consider the four nucleotides as basic states of the $(\frac{1}{2},\frac{1}{2})$ representation of the $\mathscr{U}_q(sl(2)\oplus sl(2))$ quantum enveloping algebra in the limit $q\to 0$ [1]. A triplet of nucleotides will then be obtained by constructing the tensor product of three such four-dimensional representations. Actually, this approach mimicks the group theoretical classification of baryons made out from three quarks in elementary particles physics, the building blocks being here the A, C, G, T/U nucleotides. The main and essential difference stands in the property of a codon to be an *ordered* set of three nucleotides, which is not the case for a baryon.

Constructing such pure states is made possible in the framework of any algebra $\mathscr{U}_{q\to 0}(\mathscr{G})$ with $\mathscr{G}$ being any (semi)-simple classical Lie algebra owing to the existence of a special basis, called crystal basis, in any (finite dimensional) representation of $\mathscr{G}$. The algebra $\mathscr{G}=sl(2)\oplus sl(2)$ appears the most natural for our purpose. The complementary rule in the DNA–mRNA transcription may suggest to assign a *quantum number* with opposite values to the couples (A,T/U) and (C,G). The distinction between the purine bases (A,G) and the pyrimidine ones (C,T/U) can be algebraically represented in an analogous way. Thus considering the fundamental representation $(\frac{1}{2},\frac{1}{2})$ of $sl(2)\oplus sl(2)$ and denoting $\pm$ the basis vector corresponding to the eigenvalues $\pm\frac{1}{2}$ of the J_3 generator in any of the two $sl(2)$ corresponding algebras, we will assume the following "biological" spin structure:

$$\begin{array}{ccc} & sl(2)_H & \\ C\equiv(+,+) & \longleftrightarrow & U\equiv(-,+) \\ sl(2)_V\updownarrow & & \updownarrow sl(2)_V \\ G\equiv(+,-) & \longleftrightarrow & A\equiv(-,-) \\ & sl(2)_H & \end{array} \qquad (1)$$

the subscripts H (:= horizontal) and V (:= vertical) being just added to specify the algebra.

Now, we consider the representations of $\mathscr{U}_q(sl(2))$ and more specifically the crystal bases obtained when $q\to 0$. Introducing in $\mathscr{U}_{q\to 0}(sl(2))$ the operators J_+ and J_- after modification of the corresponding simple root vectors of $\mathscr{U}_q(sl(2))$, a particular kind of basis in a $\mathscr{U}_{q\to 0}(sl(2))$-module can be defined. Such a basis is called a crystal basis and carries the property to undergo in a specially simple

way the action of the J_+ and J_- operators: as an example, for any couple of vectors u,v in the crystal basis **B**, one gets $u = J_+ v$ if and only if $v = J_- u$. More interesting for our purpose is the crystal basis in the tensorial product of two representations.

Indeed in [7] it has been shown that the tensor product of two crystal bases, labelled by J_1 and J_2, can be decomposed into a direct sum of crystal bases labelled, as in the case of the tensor product of two standard $sl(2)$ irreducible representations, by a set of integer ($J_1 + J_2$ integer) or half-integer ($J_1 + J_2$ half-integer) numbers J such that

$$|J_1 - J_2| \leq J \leq J_1 + J_2 \tag{2}$$

The new peculiar and crucial feature is that, in the limit $q \to 0$, the basis vectors of the J-space are *pure states*, that is they are the product of a state belonging to the J_1-space and of a state belonging to the J_2-space. The tensor product is not symmetric and one has to specify which is the first irreducible representation. in the tensor product, i.e. an order has to be fixed. In the framework of DNA/RNA the order appears in a very natural way as the reading frame moves toward the $5\prime \to 3\prime$ direction.

Therefore, the framework of the crystal basis model appears perfectly adapted to represent codons as composite states of the (elementary) nucleotides.

In Table 2 we report the assignments of the codons of the eukariotic code (the upper label denotes different irreducible representations) and, respectively the amino-acid content of the $\otimes^3(\frac{1}{2},\frac{1}{2})$ representations.The codon content in each of the obtained irreducible representations is also expressed at the end of this subsection.

codon	a.a	J_H J_V	$J_{H,3}$	$J_{V,3}$	codon	a.a.	J_H J_V	$J_{H,3}$	$J_{V,3}$
CCC	Pro P	3/2 3/2	3/2	3/2	UCC	Ser S	3/2 3/2	1/2	3/2
CCU	Pro P	$(1/2\ 3/2)^1$	1/2	3/2	UCU	Ser S	$(1/2\ 3/2)^1$	−1/2	3/2
CCG	Pro P	$(3/2\ 1/2)^1$	3/2	1/2	UCG	Ser S	$(3/2\ 1/2)^1$	1/2	1/2
CCA	Pro P	$(1/2\ 1/2)^1$	1/2	1/2	UCA	Ser S	$(1/2\ 1/2)^1$	−1/2	1/2
CUC	Leu L	$(1/2\ 3/2)^2$	1/2	3/2	UUC	Phe F	3/2 3/2	−1/2	3/2
CUU	Leu L	$(1/2\ 3/2)^2$	−1/2	3/2	UUU	Phe F	3/2 3/2	−3/2	3/2
CUG	Leu L	$(1/2\ 1/2)^3$	1/2	1/2	UUG	Leu L	$(3/2\ 1/2)^1$	−1/2	1/2
CUA	Leu L	$(1/2\ 1/2)^3$	−1/2	1/2	UUA	Leu L	$(3/2\ 1/2)^1$	−3/2	1/2
CGC	Arg R	$(3/2\ 1/2)^2$	3/2	1/2	UGC	Cys C	$(3/2\ 1/2)^2$	1/2	1/2
CGU	Arg R	$(1/2\ 1/2)^2$	1/2	1/2	UGU	Cys C	$(1/2\ 1/2)^2$	−1/2	1/2
CGG	Arg R	$(3/2\ 1/2)^2$	3/2	−1/2	UGG	Trp W	$(3/2\ 1/2)^2$	1/2	−1/2

Continue on the next page

codon	a.a	J_H J_V	$J_{H,3}$	$J_{V,3}$	codon	a.a.	J_H J_V	$J_{H,3}$	$J_{V,3}$
CGA	Arg R	$(1/2\ 1/2)^2$	1/2	−1/2	UGA	Ter	$(1/2\ 1/2)^2$	−1/2	−1/2
CAC	His H	$(1/2\ 1/2)^4$	1/2	1/2	UAC	Tyr Y	$(3/2\ 1/2)^2$	−1/2	1/2
CAU	His H	$(1/2\ 1/2)^4$	−1/2	1/2	UAU	Tyr Y	$(3/2\ 1/2)^2$	−3/2	1/2
CAG	Gln Q	$(1/2\ 1/2)^4$	1/2	−1/2	UAG	Ter	$(3/2\ 1/2)^2$	−1/2	−1/2
CAA	Gln Q	$(1/2\ 1/2)^4$	−1/2	−1/2	UAA	Ter	$(3/2\ 1/2)^2$	−3/2	−1/2
GCC	Ala A	3/2 3/2	3/2	1/2	ACC	Thr T	3/2 3/2	1/2	1/2
GCU	Ala A	$(1/2\ 3/2)^1$	1/2	1/2	ACU	Thr T	$(1/2\ 3/2)^1$	−1/2	1/2
GCG	Ala A	$(3/2\ 1/2)^1$	3/2	−1/2	ACG	Thr T	$(3/2\ 1/2)^1$	1/2	−1/2
GCA	Ala A	$(1/2\ 1/2)^1$	1/2	−1/2	ACA	Thr T	$(1/2\ 1/2)^1$	−1/2	−1/2
GUC	Val V	$(1/2\ 3/2)^2$	1/2	1/2	AUC	Ile I	3/2 3/2	−1/2	1/2
GUU	Val V	$(1/2\ 3/2)^2$	−1/2	1/2	AUU	Ile I	3/2 3/2	−3/2	1/2
GUG	Val V	$(1/2\ 1/2)^3$	1/2	−1/2	AUG	Met M	$(3/2\ 1/2)^1$	−1/2	−1/2
GUA	Val V	$(1/2\ 1/2)^3$	−1/2	−1/2	AUA	Ile I	$(3/2\ 1/2)^1$	−3/2	−1/2
GGC	Gly G	3/2 3/2	3/2	−1/2	AGC	Ser S	3/2 3/2	1/2	−1/2
GGU	Gly G	$(1/2\ 3/2)^1$	1/2	−1/2	AGU	Ser S	$(1/2\ 3/2)^1$	−1/2	−1/2
GGG	Gly G	3/2 3/2	3/2	−3/2	AGG	Arg R	3/2 3/2	1/2	−3/2
GGA	Gly G	$(1/2\ 3/2)^1$	1/2	−3/2	AGA	Arg R	$(1/2\ 3/2)^1$	−1/2	−3/2
GAC	Asp D	$(1/2\ 3/2)^2$	1/2	−1/2	AAC	Asn N	3/2 3/2	−1/2	−1/2
GAU	Asp D	$(1/2\ 3/2)^2$	−1/2	−1/2	AAU	Asn N	3/2 3/2	−3/2	−1/2
GAG	Glu E	$(1/2\ 3/2)^2$	1/2	−3/2	AAG	Lys K	3/2 3/2	−1/2	−3/2
GAA	Glu E	$(1/2\ 3/2)^2$	−1/2	−3/2	AAA	Lys K	3/2 3/2	−3/2	−3/2

TABLE 2: Deljenje brojeva zapisanih u potpunom komplementu

From Table 2, the dinucleotide states formed by the first two nucleotides in a codon can be put in correspondence with quadruplets, doublets or singlets of codons relative to an amino-acid. Note that the sextets (resp. triplets) are viewed as the sum of a quadruplet and a doublet (resp. a doublet and a singlet). Let us

define the "charge" Q of a dinucleotide state by

$$Q = J_{H,3} + \frac{1}{4} C_V (J_{V,3} + 1) - \frac{1}{4} \tag{3}$$

$J_{\alpha,3}$ ($\alpha = H,V$) stands for the diagonalised $sl(2)_\alpha$ generator.

The dinucleotide states are then split into two octets with respect to the charge Q: the eight *strong* dinucleotides associated to the quadruplets (as well as those included in the sextets) of codons satisfy $Q > 0$, while the eight *weak* dinucleotides associated to the doublets (as well as those included in the triplets) and eventually to the singlets of codons satisfy $Q < 0$. Let us remark that by the change $C \leftrightarrow A$ and $U \leftrightarrow G$, which is equivalent to the change of the sign of $J_{\alpha,3}$ or to reflexion with respect to the diagonals of the eq.(1), the 8 strong dinucleotides are transformed into weak ones and vice-versa.

Let us recall that a Casimir operator C_α can be defined for $\mathscr{U}_{q\to 0}(sl(2)_\alpha)$ in the crystal basis. It commutes with $J_{\alpha\pm}$ and $J_{\alpha,3}$ and its eigenvalues on any vector basis of an irreducible representation of highest weight j is $j(j+1)$, that is the same as the undeformed standard second degree Casimir operator of $sl(2)$. Its explicit expression is

$$C_\alpha = (J_{\alpha,3})^2 + \tfrac{1}{2} \sum_{n\in\mathbf{Z}_+} \sum_{k=0}^{n} (J_{\alpha-})^{n-k} (J_{\alpha+})^n (J_{\alpha-})^k \tag{4}$$

2.2. Applications

Sum rules of codon usage probabilities Let XZN be a codon in a multiplet encoding an amino acid, where the labels X,Z,N stands for any of the four bases $A,C,G,U/T$. We define the relative frequency of usage of the codon XZN as the ratio between the number of times n_{XZN} the codon XZN is used in the biosynthesis of the amino acid, and the total number n_{tot} of synthesised amino acid,. Then, the frequency of usage of a codon in a multiplet is connected, in the limit of *very large* n_{tot}, to its probability of usage $P(XZN)$:

$$P(XZN) = \lim_{n_{tot}\to\infty} \quad \frac{n_{XZN}}{n_{tot}} \tag{5}$$

with the normalization

$$P(XZA) + P(XZC) + P(XZG) + P(XZU) = 1 \tag{6}$$

The aim of the paper [2] was to investigate this aspect and to predict a general law which should be satisfied by all the biological species belonging to vertebrates.

Assuming the dependence of the amino-acid to be completely determined by the set of labels Js, we write

$$P(XZN) = P(b.s.; J_H, J_V, J_{H,3}, J_{V,3}) \tag{7}$$

Let us now make the hypothesis that we can write the r.h.s. of eq. (7) as the sum of two contributions: a universal function ρ independent on the biological species at least for vertebrates and a b.s. depending function f_{bs}, i.e.

$$P(XZN) = \rho^{XZ}(J_H, J_V, J_{H,3}, J_{V,3}) \; + \; f_{bs}^{XZ}(J_H, J_V, J_{H,3}, J_{V,3}) \tag{8}$$

¿From the analysis of the available data, we assume that the contribution of f_{bs} is not negligible but could be smaller than the one due to ρ. As each state describing a codon is labelled by the *quantum* labels of two commuting $sl(2)$, it is reasonable, at first approximation, to assume

$$f_{bs}^{XZ}(J_H, J_V, J_{H,3}, J_{V,3}) \approx F_{bs}^{XZ}(J_H; J_{H,3}) + G_{bs}^{XZ}(J_V; J_{V,3}) \tag{9}$$

Now, let us analyse in the light of the above considerations the usage probability for the quartets Ala, Gly, Pro, Thr and Val and for the quartet sub-part of the sextets Arg (i.e. the codons of the form CGN), Leu (i.e. CUN) and Ser (i.e. UCN). For Thr, Pro, Ala and Ser we can write, using Table 2 and eqs. (7)-(9), with $N = A, C, G, U$,

$$\begin{aligned} &P(NCC) + P(NCA) = \\ &\rho_{C+A}^{NC} + F_{bs}^{NC}(\tfrac{3}{2}; x) + G_{bs}^{NC}(\tfrac{3}{2}; y) + F_{bs}^{NC}(\tfrac{1}{2}; x') + G_{bs}^{NC}(\tfrac{1}{2}; y') \end{aligned} \tag{10}$$

where we have denoted by ρ_{C+A}^{NC} the sum of the contribution of the universal function (i.e. not depending on the biological species) ρ relative to NCC and NCA, while the labels x, y, x', y' depend on the nature of the first two nucleotides NC, see Table 2. For the same amino acid we can also write

$$\begin{aligned} &P(NCG) + P(NCU) = \\ &\rho_{G+U}^{NC} + F_{bs}^{NC}(\tfrac{3}{2}; x) + G_{bs}^{NC}(\tfrac{3}{2}; y) + F_{bs}^{NC}(\tfrac{1}{2}; x') + G_{bs}^{NC}(\tfrac{1}{2}; y') \end{aligned} \tag{11}$$

Using the results of Table 2, we can remark that the difference between eq. (10) and eq. (11) is a quantity independent of the biological species,

$$P(NCC) + P(NCA) - P(NCG) - P(NCU) \;=\; \rho_{C+A}^{NC} - \rho_{G+U}^{NC} \;=\; \text{Const.} \tag{12}$$

In the same way, considering the cases of Leu, Val, Arg and Gly, we obtain with $W = C, G$

$$\begin{aligned} P(WUC) + P(WUA) - P(WUG) - P(WUU) = \rho_{C+A}^{WU} - \rho_{G+U}^{WU} \;&=\; \text{Const.} \\ P(CGC) + P(CGA) - P(CGG) - P(CGU) = \rho_{C+A}^{CG} - \rho_{G+U}^{CG} \;&=\; \text{Const.} \\ P(GGC) + P(GGA) - P(GGG) - P(GGU) = \rho_{C+A}^{GG} - \rho_{G+U}^{GG} \;&=\; \text{Const.} \end{aligned} \tag{13}$$

Since the probabilities for one quadruplet are normalised to one, from eqs. (11)-(13) we deduce that for all the eight amino acids the sum of probabilities of codon usage for codons with last A and C (or U and G) nucleotide is independent of the biological species, i.e.

$$P(XZC) + P(XZA) = \text{Const.} \qquad (XZ = NC, CU, GU, CG, GG) \tag{14}$$

Moreover, assuming that for sextets the functions F and G depend really on the nature of the encoded amino acid rather than on the dinucleotide, we derive in a completely analogous way as above that for the amino acid Ser the sum $P'_{C+A}(S) = P(UCA) + P(AGC)$ is independent of the biological species. Note that

we normalize to 1 the probabilities of a quartet in a sextet. The results are reported in Table 3.

A statistical discussion of the sum rules, in the more general context of correlations between the probabilities $P(XZN)$, can be found in [9].

An analysis with more recent data for more biological species can be found in [10].

2.3. Physico-chemical properties of amino-acids: relations and predictions

It is a known observation that a relationship exists between the codons and the physical-chemical properties of the coded amino acids. The observed pattern is read either as a relic of some kind of interaction between the amino acids and the nucleotides at an early stage of evolution or as the existence of a mechanism relating the properties of codons with those of amino acids.

It is also observed that the relationship depends essentially on the nature of the *second* nucleotide in the codons and it holds when the second nucleotide is A, U, C, not when it is G. To our knowledge neither the anomalous behaviour of G nor the existence of a closest relationship between some of the amino acids is understood. In [3] we provided an explanation of both these facts in the framework of the crystal basis model of the genetic code.

Relationship between the physical-chemical properties of amino acids We assume that some physical-chemical property of a given amino acid are related to the nature of the codons, in particular they depend on the following mathematical features, written in hierarchical order:

1. the irreducible representation of the dinucleotide formed by the first two nucleotides;
2. the sign of the charge Q eq.(3) on the dinucleotide state;
3. the value of the third component of $J_{V,3}$ inside a fixed irreducible representations for the dinucleotides;
4. the upper label(s) of the codon irreducible representation(s);

Not all the physical-chemical properties are supposed to follow the scheme above; some of them are essentially given by the specific chemical structure of the amino acid itself.

We have compared our theoretical predictions with 10 physical-chemical properties:

- the Chou-Fasman conformational parameters P_α, P_β and P_τ which gives a measure of the probability of the amino acids to form respectively a helix, a sheet and a turn. The sum $P_\alpha + P_\beta$ appears more appropriate to characterise the generic structure forming potential and the difference $P_\alpha - P_\beta$ the helix forming potential, this quantity depending more on the particular amino acid. So we compare with $P_\alpha + P_\beta$ and P_τ;
- the Grantham polarity P_G;
- the relative hydrophilicity R_f;

- the thermodynamic activation parameters at 298 K: ΔH (enthalpy, in kJ/mol), ΔG (free energy, in kJ/mol) and ΔS (entropy, in J/mole/K);
- the negative of the logarithm of the dissociation constants at 298 K: pK_a for the α-COOH group and pK_b for the α-NH_3^+ group;
- the isoelectronic point pI, i.e. the pH value at which no electrophoresis occurs.

In summary, for a detailed discussion see [3], we found that the values of physical-chemical properties show, with a few exceptions, a pattern of correlations which is expected from the assumptions of the crystal basis model. The remarked property that the amino acids coded by codons whose second nucleotide is G do not share similarity in the physical-chemical properties with other amino acids does find an explication in the model, as it is immediate to verify that there are no two states with G in second position which share simultaneously the properties of belonging to the same irreducible representation and being characterised by the same value of Q.

Finally, some quantities not yet measured at the time we completed the work, are predicted: for Asp and Glu, one should find $\Delta H \approx 60$ kJ/mol, $-\Delta S \approx 135$ kJ/mol/K and $\Delta G \approx 100$ kJ/mol.

3. PART 2: A "minimum" principle in the genetic code

3.1. A "minimum" principle in the mRNA editing

The mathematical formulation of a sequence in RNA or DNA in the crystal basis model allows to investigate if some "minimum" principle can be applied to the genetic code.

In [8], we have investigated the possibility to explain the position of a nucleotide insertion in mRNA, the so called mRNA editing. The deep mechanism which causes RNA editing is still unknown. The understanding of the event is complicated: from a thermodynamics point of view a change, i.e. C $\rightarrow$ U, takes place if it is favored in the change of entalpy or entropy, but should this be the case, the change should appear in all the organisms. Moreover from a microscopic (quantum mechanical) point of view, the change should occur in both directions, i.e.. C $\leftrightarrow$ U. It seems that the primary aim of mRNA editing is the evolution and conservation of protein structures, creating a meaningful coding sequence specific for a particular amino acid sequence.

The purpose of the paper [8] was to propose an effective model to describe the RNA editing. Our model does not explain why, where and in which organisms editing happens, but it gives a framework to understand some specific features of the phenomenon.

A consequence of the crystal basis model is that any nucleotide sequence is characterized as an element of a vector space. Therefore, functions can be defined on this space and can be computed on the sequence of codons. In particular any codon is identified by a set of four half-integer labels and functions can be defined on the codons. We make the assumption that the location sites for the

TABLE 3: Mean value, standard deviation and χ^2 for the probabilities P_N corresponding to quartets, computed over 35 vertebrates.

Pro	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.325	0.279	0.271	0.124	0.596	0.605	0.450
σ	0.046	0.034	0.037	0.028	0.029	0.037	0.057
χ^2	6.4	9.7	9.7	5.1	11.9	4.3	5.2
Thr	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.373	0.235	0.269	0.123	0.642	0.608	0.496
σ	0.053	0.031	0.042	0.027	0.027	0.034	0.066
χ^2	4.8	10.2	7.6	9.9	6.4	7.1	12.9
Ala	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.390	0.282	0.216	0.112	0.605	0.672	0.502
σ	0.050	0.035	0.038	0.027	0.034	0.033	0.060
χ^2	6.6	4.8	3.4	1.3	11.6	3.0	7.8
Ser	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.373	0.307	0.224	0.096	0.597	0.680	0.469
σ	0.039	0.030	0.036	0.018	0.021	0.037	0.047
χ^2	8.7	19.5	5.7	5.3	1.0	5.4	12.3
Val	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.259	0.173	0.101	0.466	0.361	0.432	0.725
σ	0.026	0.035	0.033	0.045	0.027	0.032	0.057
χ^2	4.8	17.9	9.7	12.9	10.6	4.5	6.2
Leu	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.252	0.150	0.083	0.516	0.334	0.402	0.767
σ	0.019	0.034	0.027	0.044	0.022	0.024	0.056
χ^2	8.7	5.1	10.8	13.6	2.2	7.9	13.8
Arg	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.343	0.181	0.184	0.292	0.527	0.524	0.635
σ	0.039	0.061	0.031	0.056	0.024	0.068	0.064
χ^2	7.3	16.2	10.5	14.6	3.5	15.8	12.7
Gly	P_C	P_U	P_A	P_G	P_{C+A}	P_{C+U}	P_{C+G}
$\bar{x}$	0.321	0.185	0.271	0.223	0.592	0.506	0.544
σ	0.041	0.041	0.042	0.047	0.020	0.025	0.079
χ^2	5.5	15.3	14.8	18.8	9.9	10.7	19.7

insertion of a nucleotide should minimize the following function for the mRNA or cDNA

$$\mathscr{A}_0 = \exp\left[-\sum_k 4\alpha_c C_H^k + 4\beta_c C_V^k + 2\gamma_c J_{3,H}^k\right] \tag{15}$$

where the sum in k is over all the codons in the edited sequence, C_H^k (C_V^k) and $J_{3,H}^k$ ($J_{3,V}^k$), are the values of the *Casimir* operator, see eq.(4) and of the third component of the generator of the $sl(2)_H$ ($sl(2)_V$), in the irreducible representation to which the k-th codon belongs, see Table 2. In (15) the simplified assumption that the dependence of $\mathscr{A}_0$ on the irreducible representation to which the codon

belongs is given only by the values of the Casimir operators has been made. The parameters $\alpha_c, \beta_c, \gamma_c$ are constants, depending on the biological species.

The minimum of $\mathscr{A}_0$ has to be computed in the whole set of configurations satisfying to the constraints:

- the starting point should be the mtDNA;
- the final peptide chain should not be modified.

It is obvious that the global minimization of expression eq.(15) is ensured if $\mathscr{A}_0$ takes the smallest value locally, i.e. in the neighborhood of each insertion site. The form of the function $\mathscr{A}_0$ is rather arbitrary; one of the reasons of this choice is that the chosen expression is computationally quite easily tractable. If the parameters $\alpha_c, \beta_c, \gamma_c$ are strictly positive with $\gamma_c/6 > \beta_c > \alpha_c$, the minimization of eq.(15) explains the observed configurations in almost all the considered cases, for more details see [8].

In order to take into account the observed fact that the dinucleotide preceding the insertion site is predominantly a purine-pyrimidine, in some cases eq.(15) has to be multiplied by $\mathscr{A}_1$

$$\mathscr{A}_1 = \exp\left[\sum_i -4\,\omega_{1c}\, j^{(i)}_{3,V} \cdot j^{(i-1)}_{3,V} + 4\,\omega_{2c}\, j^{(i)}_{3,V} \cdot j^{(i-2)}_{3,V}\right] \tag{16}$$

The sum in i is over the insertion sites and $j^{(i-n)}_{3,V}$ is the value of the third component of the generator of V-$sl(2)$ of the n-th nucleotide preceding the inserted nucleotide C (i.e. $+1/2$ for C, U and $-1/2$ for G, A) and ω_{1c}, ω_{2c} are constants, depending on the biological species.

Moreover, we make the assumption that the location sites for the insertion of a U nuleotide should minimize the following function for the mRNA:

$$\mathscr{A}'_0 = \exp\left[-\sum_k 4\alpha_u C^k_H + 4\beta_u C^k_V + 2\gamma_u J^k_{3,H}\right] \tag{17}$$

We have shown:

- for Physarum polycephalum, in 110 of the 114 sites in which there is an insertion of C or U, and in all the cases where also an insertion of purine can produce the same amino acid, the observed mRNA editing makes use of the nucleotide C or U which does minimize $\mathscr{A} = \mathscr{A}_0\mathscr{A}_1$;
- for kinetoplastid protozoa genes, the U insertion in all the cases, but two, the function $\mathscr{A}'$ is minimized. This last function is also minimized in the case of C $\rightarrow$ U substitution editing.

3.2. The "minimum" principle to explain the codon bias

A codon-anticodon interaction potential has beeb proposed [4], still in the framework of the "Crystal Basis Model". The minimization of this interaction has allowed to determine:

- the structure of the minimum set of 22 anticodons allowing the translational-transcription for animal mitochondrial code [4],
- to study the evolution of the genetic code, with 20 amino-acids encoded from the beginning, in particular the determination of the structure of the anticodons in the Ancient, Archetypal and Early Genetic codes was obtained [5]. Most of our results agree with the generally accepted scheme;
- to provide inequalities between codon usage probabilities for quartets of codons [6].

In the following we will review the third point, for a review of the first two points see [11].

As already stated the genetic code is degenerate in the sense that a multiplet is used to encode most of the amino-acids. Some codons in the multiplets are used much more frequently than others to encode a particular amino-acid, i.e. there is a "codon usage bias". The non-uniform usage of synonymous codons is a widespread phenomenon and it is experimentally observed that the pattern of codon usage varies between species.

No clear indication comes out for the existence of one or more factors which universally engender the codon bias, on the contrary the role of some factors is controversial.

In [4] we have described the codon-anticodon interaction by means of an operator $\mathscr{T}$ given by

$$\mathscr{T} = 8c_H J_H^c \cdot J_H^a + 8c_V J_V^c \cdot J_V^a \tag{18}$$

which has to be computed between the "states", which can be read from table describing the codon and anticodon in the "crystal basis model", where:

- $c_H.c_V$ are constants depending on the "biological species" and weakly depending on the encoded a.a., as we will later specify.
- J_H^c, J_V^c (resp. J_H^a, J_V^a) are the labels of $\mathscr{U}_{q\to 0}(su(2)_H \oplus su(2)_V)$ specifying the state
  describing the codon XYZ (resp. the anticodon NY_cX_c pairing the codon XYZ).
- $J_\alpha^c \cdot J_\alpha^a$ ($\alpha = H,V$) should be read as

$$J_\alpha^c \cdot J_\alpha^a = \frac{1}{2}\left\{ (J_\alpha^c \oplus {J_\alpha}^a)^2 - (J_\alpha^c)^2 - (J_\alpha^a)^2 \right\} \tag{19}$$

  and $J_\alpha^c \oplus J_\alpha^a \equiv J_\alpha^T$ stands for the irreducible representation which the codon-anticodon state under consideration belongs to, the tensor product of J_α^c and J_α^a being performed according to the rule of [7], choosing the codon as first vector and the anticodon as second vector. Note that ${J_\alpha}^2$ should be read as the Casimir operator whose eigenvalues are given by $J_\alpha(J_\alpha+1)$.

We write both codons (c) and anticodons (a) in 5" $\to$ 3" direction. As an anticodon is antiparallel to codon, the 1st nucleotide (respectively the 3rd nucleotide) of the anticodon is paired to the 3rd (respectively the 1st) nucleotide of the codon.

In the following, we will be concerned about amino acids encoded by quartets. For the ones encoded by a sextet, that we consider as the sum of a quartet

and a doublet, only the quartet will be considered. The method we developed is essentially based on the determination of the minimum values of an operator which can be seen as an interaction potential between a codon and its corresponding anticodon. A possible general pattern of the bias is searched by deriving inequalities for the codon usage probabilities.

We have to minimize an expression of the type:

$$\begin{aligned}
&\mathscr{T}_{av}(N'YX'',XYN) = \sum_N P_N \; < N'Y''X''|\mathscr{T}|XYN > \\
&= P_C \; < N'Y''X''|\mathscr{T}|XYC > + P_U \; < N'Y''X''|\mathscr{T}|XYU > \\
&+ P_G \; < N'Y''X''|\mathscr{T}|XYG > + P_A \; < N'Y''X''|\mathscr{T}|XYA >
\end{aligned} \tag{20}$$

Let us recall that the expression $< N'Y''X''|\mathscr{T}|XYN >$ has to be read as

$$\begin{aligned}
&< N'Y''X''|\mathscr{T}|XYN > \equiv \mathscr{T}\left(|XYN > \otimes |N'Y''X'' >\right) \\
&= \mathscr{T}\left(|J^c_H, J^c_V; J^c_{H,3}, J^c_{V,3} > \otimes |J^a_H, J^a_V; J^a_{H,3}, J^a_{V,3} >\right) \\
&= \lambda\left(|J^c_H, J^c_V; J^c_{H,3}, J^c_{V,3} > \otimes |J^a_H, J^a_V; J^a_{H,3}, J^a_{V,3} >\right)
\end{aligned} \tag{21}$$

where we have used the correspondence

$$\begin{aligned}
|XYN > \; &\rightarrow |J^c_H, J^c_V; J^c_{H,3}, J^c_{V,3} > \\
|N'Y''X'' > \; &\rightarrow |J^a_H, J^a_V; J^a_{H,3}, J^a_{V,3} >
\end{aligned} \tag{22}$$

and λ is the eigenvalue of $\mathscr{T}$ on the state $|J^c_H, J^c_V; J^c_{H,3}, J^c_{V,3} > \otimes |J^a_H, J^a_V; J^a_{H,3}, J^a_{V,3} >$, see [4] for more details. As the P_{XYN} have to satisfy, in addition to eq.(6), a set of unknown constraints, we cannot impose the minimization condition in a rigorous manner, so we proceed by a heuristic method. Using the results of Subsection 2.2, we are left with only two probabilities in eq.(20) and we try to argue which from the two present P_{XYN} is enhanced respect to the other one. For this aim we compare the two probabilities which appear after the substitution of the other two, using eq.(14), that have the greatest coefficient. In this way we will get in our expression a constant terms, depending on c_H, c_V and, generally, on K (K being the constant which appears in the r.h.s. of eq.(14), which has the highest possible value, without, possibly, any specific assumption on the value of the parameters, except for the assumed sign. Then, in order to minimize the expression, it is reasonable to require that the probability with the lowest coefficient has a higher value than the other one. Nextly, in some cases, we can derive another inequality for the complementary probabilities, according to $K > 0,5$ or $K < 0,5$.

For a more detailed discussion of the difference between the minimization procedure for the Early Genetic Code and the Eukaryotic Genetic Code, as well as on the assumed behavior of the coefficient c_H and c_V we refer to [6].

The outcomes derived, which we summarize in Table 4, are in an amazing agreement with the observed data, nevertheless the over-simplifying assumptions of our theoretical scheme and despite that in the real world the number of operating anticodons is greater that the minimum number 31, which implies that the matching of an a.a. encoded by a quartet is done by more than two

anticodons. Moreover let us remark that the results found in the Early Genetic Code survive in the Eukaryotic Genetic Code, suggesting that we have caught some feature of a very relevant mechanism. So we argue that codon-anticodon interaction plays a relevant role in the codon usage bias. Moreover it seems that, in despite of its apparently fragmented behavior, the codon bias exhibits a sort of universal feature that our approach and the Crystal Basis Model is able to take into account.

TABLE 4: Inequalities derived in the Early and in the Eukaryotic Genetic Code. The value of the parameters c_H and c_V is different in the two codes.

	Early Code		**Eukaryotic Code**	
a.a.	Parameters	Inequalities	Parameters	Inequalities
Thr	—	$P_C > P_G$	$\|c_H\| < 3\|c_V\|$ $\|c_H\| > 3\|c_V\|$	$P_C > P_U$ $P_C > P_G$
Arg	—	—	$\|c_V\| < \|c_H\| < 2\|c_V\|$ $\|c_H\| < \|c_V\|$	$P_C > P_G$ $P_C > P_U$
Pro	$\|c_V\| < c_H$ $\|c_V\| > c_H$	$P_A > P_U$ $P_A > P_G$	$\|c_V\| < 1/4\|c_H\|$ $\|c_V\| > 1/4\|c_H\|$	$P_C > P_U$ $P_U > P_C$,$P_A > P_G$
Leu	—	$P_G > P_C$	$\|c_H\| < 2/3\|c_V\|$, $\|c_H\| > 2/3\|c_V\|$,	$P_U > P_C$ $P_C > P_U$, $P_G > P_A$
Ala	$\|c_V\| < c_H$ $\|c_V\| > c_H$	$P_U > P_A$, $P_C > P_G$ $P_U > P_C$, $P_A > P_G$	—	$P_C > P_U$
Gly	—	—	$\|c_H\| < \|c_V\|$ $\|c_H\| > \|c_V\|$	$P_C > P_G$ $P_G > P_C$, $P_A > P_U$
Val	$\|c_V\| < 4c_H$ $\|c_V\| > 4c_H$	$P_C > P_G$,$P_U > P_A$ $P_C > P_U$, $P_G > P_A$	$\|c_H\| < 3\|c_V\|$ $\|c_H\| > 3\|c_V\|$	$P_C > P_U$,$P_G > P_A$ $P_C > P_G$, $P_U > P_A$
Ser	$\|c_V\| < 3c_H$ $\|c_V\| > 3c_H$	$P_G > P_C$,$P_A > P_U$ $P_G > P_A$,$P_C > P_U$	$\|c_V\| < \frac{5}{4}\|c_H\|$ $\|c_V\| > \frac{5}{4}\|c_H\|$	$P_C > P_U$ $P_U > P_C$, $P_A > P_G$

4. Conclusions

The presented model for the genetic code is an attempt towards a theoretical and mathematical approach in the complex domain of the sciences of life.

Starting from the idea that codons can be seen as composite states of the 4 nucleotides, we have built up a mathematical parametrization which can be useful to deal with several and different aspect of the genetic code as well as of the DNA or RNA sequences.

¿From the so far obtained results, we believe worthwhile to pursue investigations in the framework of our model. Such developments could be carried out as well as in the mathematical side as in the phenomenological one. Let

us for example note the construction of a distance [3] between two sequences of DNA or RNA [13] still in the framework of our model: this work deserves to be developed and applied.

5. Acknowledgments

P.S. would like to express his gratitude to Professors Gordana Pavlovic-Lazetic, Branko Dragovich and Nenad Mitic for their kind invitation to present our results at the BelBI2016 International Symposium held at the University of Belgrade, Serbia, and for contributing so efficiently to the success of this very interesting conference. He is specially indebted to Professor Dragovich for his constant and friendly support in the development of our model.

References

1. L. Frappat, A. Sciarrino and P. Sorba, *Phys.Lett. A* **250** 214, (1998).
2. L. Frappat, A. Sciarrino and P. Sorba, *Phys.Lett. A* **311** 264, (2003).
3. L. Frappat, A. Sciarrino and P. Sorba, *J.Biol.Phys.* **28** 17, (2002).
4. A. Sciarrino and P. Sorba, *BioSystems* **107** 113, (2012).
5. A. Sciarrino and P. Sorba, *BioSystems* **111** 175, (2013).
6. A. Sciarrino and P. Sorba,*BioSystems* **141** 20, (2016).
7. M. Kashiwara, *Commun.Math.Phys.* **133**, 249 (1990).
8. L. Frappat, A. Sciarrino and P. Sorba, *J.Biol.Phys.* **28** 27, (2002).
9. L. Frappat, A. Sciarrino and P. Sorba, *Physica A* **351** 461, (2005).
10. D. Cocurullo and A. Sciarrino, *arXiv* **1609.02141v1** (2016).
11. A. Sciarrino and P. Sorba, *p-Adic Numbers, Ultrametric Analysis and Applications* **6** 257, (2014).
12. B.Dragovich and A.Dragovich, *arXiv* **0707.3043v1**, *The Computer Journal* **53(4)** 432, (2010).
13. A. Sciarrino, *arXiv* **1703.00445v1** (2017).

[3] Note that in a different context, but still in the framework of the genetic code, another type of metric, namely an ultrametric one, is proposed to determine a (p-adic) distance between codons: cf the talk of B.Dragovich at this meeting and [12] .

Algebraic topology of multi-brain graphs: Methods to study the social impact and other factors onto functional brain connections

Bosiljka Tadić[1] and Miroslav Andjelković[2]

[1] Department of Theoretical Physics, Jožef Stefan Institute, Jamova 39, Ljubljana, Slovenia
bosiljka.tadic@ijs.si
[2] Institute for Nuclear Science, Vinča, Belgrade, Serbia
mandjelkovic@vin.bg.ac.rs

Abstract. We discuss the perspectives of the algebraic topology of graphs in the analysis of functional multi-brain networks, based on the original work in [1]. The multi-brain graphs represent the correlations among the sets of EEG signals simultaneously recorded during the speaker-listener communications. We demonstrate how the analysis reveals the structure of the higher-order complexes in the active brain areas of each speaker and listener as well as the composition of the cross-brain correlations. Further, we discuss the potentials of the approach to study the temporal development of attention and detecting the disease-related shifts in the brain activity patterns from the suitably recorded brain imaging data.
Keywords: brain imaging data, algebraic topology of graphs, brain functional networks, inter-brain coordination

1. Introduction

In recent years, mapping the brain imaging data onto brain networks and the objective analysis using graph theory methods provided a new framework for better understanding the functional brain connections. In particular, it has been recognised which brain areas are related to information processing, cognitive control, the perception of space, time, numbers, and languages, the impact of emotions, substances, or the presence of disease [2–6]. On the other hand, the architecture of brain connections underlying human social behaviour, known as the social brain [7, 8], remains largely unexplored. The reasons are twofold. First, it is technically demanding to provide a simultaneous recording of multiple persons during the social interactions. Recently, brain imaging methods are being extended to simultaneous recording the brain activity in a group of individuals during a social communication [9–11]. The data collected in these studies provide a basis to analyse the social impact onto the brain function. Moreover, complex interactions among many different brain areas underly the social conduct and can be influenced by the social activity. The mechanisms include the mirror neurones, synchronization among particular brain regions, the use of emotions, as well as the complex processing and interpretation of

different aspects of the communicated contents and stimuli [7–11]. Hence, it is the *organisation of connections into different higher-order structures* that can potentially measure the functional patterns of the social brain rather than the mere presence of a particular relationship. These structural characteristics of the brain connectivity go beyond the standard graph theory methods, which are commonly applied to analyse the brain networks [2, 3]. Moreover, the corresponding activity patterns are likely to involve the nodes belonging to different standard modules, detectable at the mesoscopic scale [12].

To analyse the higher order structures occurring in the functional muti-brain networks, recently, we proposed [1] to apply the methods of the algebraic topology of graphs [13–18]. In particular, we consider the correlations among sets of EEG signals, which are simultaneously recorded in a group of participants during the speakerlistener communications [11]. We have demonstrated that, despite the similarity in the brain connectivity patterns, each brain possesses specific connectivity that gives raise the higher organised structures. These higher-order complexes, readily detectable by mathematical techniques of Q-analysis [18], are strictly related to the brain activity level and positively correlate with the participant's self-rating of the understanding the communicated subject and the narrative qualities of the speaker. Hence, the techniques of the algebraic topology open a new perspective towards understanding the social impact to the brain functional connections, a subject that remains largely elusive to the standard methods.

Here, we briefly summarise the methodology and demonstrate how the introduced topology measures correlate with the brain activity of the individual and cross-brain connections. As the example, we consider the data of two participants with a different role. Further, we discuss the potentials of the approach to analyse the shifts in the brain activity in the course of a process and possibility to detect a pattern related to a particular disease.

2. Multi-Brain Graphs and their Hidden Topology

The considered EEG signals recorded in [11] comprise of 63 channels in each of 14 individuals; the correlation matrix thus consists of $882\times$ 882 array elements, representing the Pearson's coefficient of all pairs of the channels, filtered to remove the redundant correlations and thresholded to keep the correlations where a non-Gaussian distribution applies (see detailed discussion in [1]). The resulting multi-brain graph is then analysed using the methods of graph theory and algebraic topology of graphs.

In the entire multi-brain network (MBN) we distinguish the single-brain networks (SBNs), constituting the 14 diagonal blocks of the size 63×63, the cross-linking between each pair of brains, identified as off-diagonal blocks 63×63 elements, as well as the mesoscopic structures (communities) extending over many brains. In general, the correlations inside a particular SBN are stronger than the inter-brain connections. However, a significant number of inter-brain connection remain after the filtering procedure, suggesting the substantially important connections between different brains. The analysis of the MBN and its subgraphs

above reveals the structural characteristics described in the following, cf. Figs. 1, 2, 3, 4.

– The community structure beyond SBNs involves parts of different brains, which appear to be connected through their frontal-to-frontal or parietal-to-parietal areas. Fig. 1 shows several such communities in the MBN evaluated at a larger correlation threshold as compared to the case discussed in [1]. Notably, the density of the inter-brain connections between frontal brain areas is large leading to topologically recognised communities. Similarly, the communities of parietal brain areas belonging to different brains can be identified. While, the frontal-parietal connections within the same brain, although being stronger, are rare and the corresponding brain areas belong to two distinct communities.

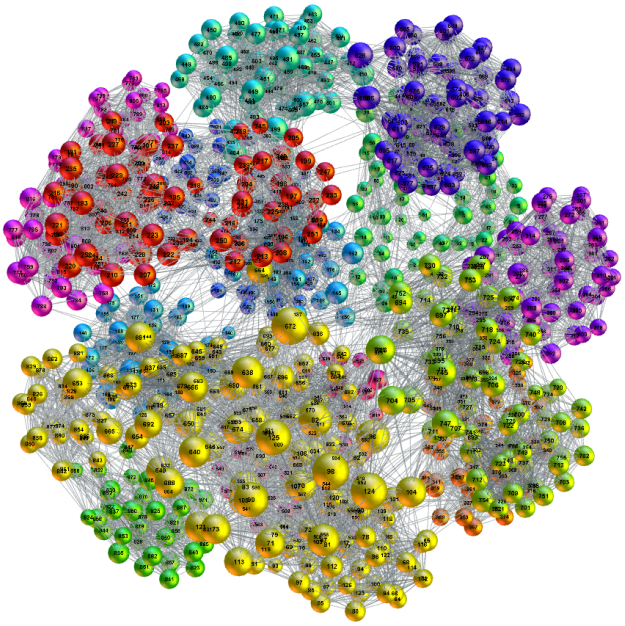


FIG. 1: Multi-brain network at elevated threshold of the correlations. The communities, marked by different colors, occur across brains by connecting frontal-to-frontal and parietal-to-parietal brain areas.

– While the SBN are statistically similar at the graph level, i.e., by the number of nodes and links, they differ by the presence/absence of a particular link (network distances) and even more significantly by the organisation of these links. The higher-order organised structures are recognised by the presence of simplexes (triangles, tetrahedra, and cliques of potentially higher dimension) and their complexes, identifiable by the appropriate algorithms [17]. We have shown [1] that the differences in the topology of SBN vary with the

quality of the speakerlistener communication, which also depends on the communicated content.

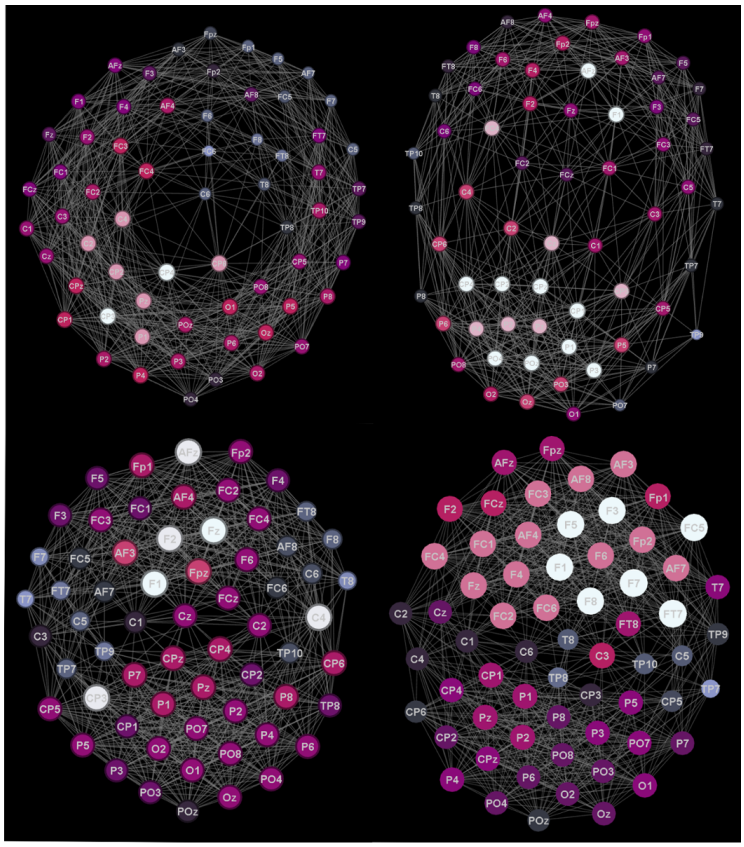


FIG. 2: The brain activity networks corresponding to the correlations of the EEG signals, indicated by the labels, for two speakers (top row), and two listeners (bottom row).

– The inter-brain connections, for instance, between the speaker and the listener shown in Fig. 3, are closely related to the quality of coordination between the two brains. They reflect both the similarities/differences in the brain activity patterns, which can be detected by comparisons of the corresponding SBNs, as well as the high/low coordination between the brains. As discussed in detail in [1], the inter-brain coordination which is represented by a significant number of cross-brain links and the presence of the higher organised structures corresponds to a right understanding of the communicated subject and the high grades to the speaker's narrative qualities. On the other hand, the listener's self-rating experience which includes the low narrative attributes of the speaker and uninteresting or confusing subject is

also manifested in the presence a few inter-brain links and poor topology of the cross-brain graph, cf. Fig. 3 left panel.

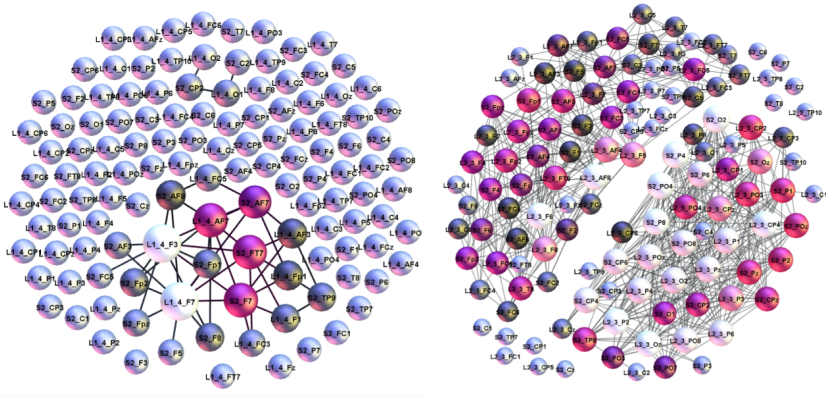


FIG. 3: Cross-brain connections corresponding to a weak speaker–listener coordination (left) and a good coordination (right).

Clique-complex algorithms determine the appearance of cliques in a particular graph [18, 17]; the presence and the organisation of these cliques are characterised by the structure vectors. In particular, the components Q_q of the first structure vector (FSV) represent the number of the q-connected classes, where $q = 0, 1, 2, \cdots q_{max}$ indicate the topology levels and q_{max} coincides with the rank of the largest cliques in the graph. The second structure vector (SSV) components n_q represent the number of cliques of the order q and larger. The elements of the third structure vector (TSV) are then derived as $\hat{Q}_q = 1 = Q_q/n_q$; they represent a useful measure of the connectivity among the cliques at the indicated topology level q. The Q-analysis of the SBNs reveals the organisation of the active areas in the speakers and listeners brain. Further, applying a similar analysis to the off-diagonal blocks, we determine the composition of the cross-brain correlations.

Fig. 4 in the left panel displays the FSV of the single-brain networks of the two speakers and the two groups of six participants. It is remarkable that the listener's brain activity patterns systematically exhibit a higher topological complexity than the speaker's. Also, compared to the SBN in Fig. 2, the majority of the listeners exhibit a larger similarity to the speaker2 than the speaker1. In [1], the distances between the brain patterns, which is evaluated by comparing the presence of the exact link in two brains, suggest that all listeners are much closer to the speaker2 than to the speaker1. Interestingly, the listener's rating available from the empirical data indicates the low narrating quality and attractiveness of the speaker1 as well as the poor understanding of the narrated subject. Similarly, the cross-brain graph of a listener with the speaker1 exhibits much simpler structure than the one with the speaker2. These graphs are displayed in Fig. 3 and the corresponding TSV in Fig. 4 right panel.

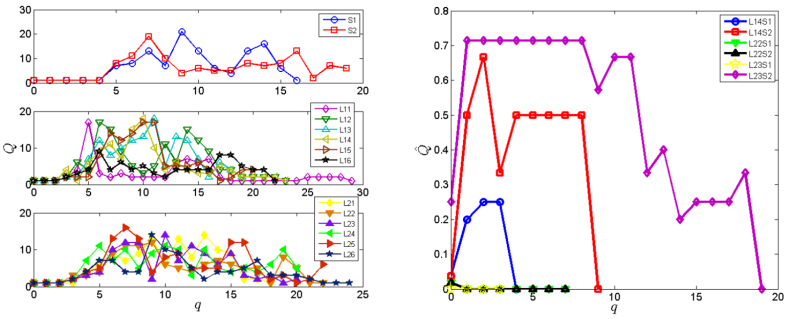


FIG. 4: (left) The components of the first structure vector Q_q as a function of the topology level q computed for the SBN for the two groups of six listeners and the two speakers. (right) The third structure vector of the cross-brain graphs shown in Fig. 3.

3. Discussion: further applications and potentials of the methodology

The algebraic topology techniques described in the mathematical literature [13–18] have been applied to describe the structure and dynamics of various complex systems. For instance, the structure of simplexes was shown to suitably characterises the hierarchical organisation of the online social networks [19], the characterization and design of the functional materials [20], and the changes of the dynamical regimes near the jamming transition [21]. In the context of the functional brain networks [1, 22], the methodology can capture the subtle shifts in the topology that underly the differences in the brain functional patterns. The structural complexity of these brain networks is quantified by the number of simplexes and the structure of the shared faces at each topological level. Our analysis of multi-brain networks and their characteristic subgraphs (single-brain networks, cross-brain graphs, communities) suggests that the topology analysis can suitably capture the inter-brain coordinations.

In this context, the applications of the algebraic topology methods to analyse various brain graphs constructed from the brain imaging data is still at the beginning. It should be stressed that, while the EEG signals are suitable to embody temporally relevant fluctuations in the brain activity, they are measured at the scalp. Therefore, often the inverse algorithms are needed to evaluate the deeper cortical sources of the signal and to reconstruct their spatial fluctuations [23]. Nevertheless, given the synchronous brain dynamics, the EEG recordings are nowadays considered as valuable sources of the overall brain activity [2]. The approach described in [1] can be directly applied to analyse the social impact onto the brain functional pattern using the simultaneous brain imaging in various situations. Furthermore, our results indicate that in each SBN topology, the changes can be detected over time and related to the variations in the attention and the potential role of emotions and other factors in the perception

of the communicated subject [8, 4]. As an important future application, we see the possibility to detect the changes in the brain activity patterns that might be caused by the presence of a neuroactive substance or a particular disease.

References

1. Tadić, B., Andjelković, M., Boškoska, B.M., and Levnajić, Z.: Algebraic Topology of Multi-Brain Connectivity Networks Reveals Dissimilarity in Functional Patterns during Spoken Communications. PLoS ONE 11(11): e0166787 (2016)
2. Sporns, O.: Structure and function of complex brain networks. Dialogues Clin. Neurosci., 15, No. 3, 247–262, (2013)
3. De Vico Falani, F., Richiardi, J., Chavez, M., Achard, S.: Graph analysis of functional brain networks: practical issues in translational neuroscience, arXiv:1406.7931
4. Garcia-Martinez B, Martinez-Rodrigo A, Zangrniz Cantabrana R, Pastor Garcia JM, Alcaraz R. Application of Entropy-Based Metrics to Identify Emotional Distress from Electroencephalographic Recordings. Entropy (2016) 18(6), 221. Available from: http://www.mdpi.com/1099-4300/18/6/221. doi: 10.3390/e18060221
5. Padovani EC. Characterisation of the Community Structure of Large-Scale Functional Brain Networks During Ketamine-Mdetomidine Anesthetic Induction. arxiv:q-bio (2016) arXiv:1604.00002. Available from: https://arxiv.org/abs/1606.04719.
6. Zeng L.-L. et al.: Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis, Brain, Vol.? (2012)
7. Adolphs R. The Social Brain: Neural Basis of Social Knowledge. Annu Rev Psychol (2009) 60, 693716. doi: 10.1146/annurev.psych.60.110707.163514. pmid:18771388
8. Krauzlis RJ, Bollimunta A, Arcizet F, Wang L. Attention as an effect not a cause. Trends in Cognitive Sciences (2016) 18(9), 457464. doi: 10.1016/j.tics.2014.05.008.
9. Yun K, Watanabe K, Shimojo S. Interpersonal body and neural synchronization as a marker of implicit social interaction. Sci Rep (2012) 2, 959. doi: 10.1038/srep00959. pmid:23233878
10. Duan L, Liu WJ, Dai RN, Li R, Lu CM, Huang YX, Zhu CZ. Cross-Brain Neurofeedback: Scientific Concept and Experimental Platform. PLoS ONE (2013) 8(5), e64590. doi: 10.1371/journal.pone.0064590. pmid:23691253
11. Kuhlen, A.K. et al.: Content-specific coordination of listeners to speakers EEG during communication, Frontiers Human Neurosci., Vol. 6, 266 (2012)
12. Gronchi G, Guazzini A, Massaro E, Bagnoli F. Mapping cortical functions with a local community detection algorithm. Journal of Complex Networks (2014) 2, 637653. doi: 10.1093/comnet/cnu035.
13. Jonsson J. Simplicial Complexes of Graphs. Lecture Notes in Mathematics, Springer-Verlag, Berlin; 2008.
14. Atkin, R.H.: An algebra for patterns on a complex, II. International Journal of Man-Machine Studies.8(5):483 (1976) Available from: http://www.sciencedirect.com/science/article/pii/S0020737376800156
15. Kozlov, D.: Combinatorial Algebraic Topology. Springer Series "Algorithms and Computation in Mathematics", Vol. 21, Springer-Verlag Berlin Heidelberg (2008)
16. Bandelt, H.J., and Chepoi, V.: Metric graph theory and geometry: a survey, in Goodman, J. E.; Pach, J.; Pollack, R., Eds. "Surveys on Discrete and Computational Geometry: Twenty Years Later". vol. 453. Providence, RI: AMS (2008)
17. Bron, C., Kerbosch, J.: Finding all cliques of an undirected graph. Comm ACM 16 (1973)

18. Gould, P.: Q-analysis, or a language of structure: an introduction for social scientists, geographers and planners. International Journal of Man-Machine Studies.13(2):169 (1980) Available from: http://www.sciencedirect.com/science/article/pii/S0020737380800095.
19. Andjelković M., Tadić B., Maletić, S., Rajkovć, M.: Hierarchical sequencing of online social graphs. Physica A: Statistical Mechanics and its Applications. 436:582 (2015). Available from: http://www.sciencedirect.com/science/article/pii/S0378437115004902.
20. Kotani, M., and Ikeda, S.: Materials inspired by mathematics, Sci Technol Adv mater. 17(1), 253-259 (2016)
21. Andjelković, M., Gupte, N., and Tadć, B.: Hidden geometry of traffic jamming. Phys Rev E 91:052817 (2015). Available from: http://link.aps.org/doi/10.1103/PhysRevE.91.052817.
22. Andjelković, M., and Tadić, B.: Algebraic topology of multi-brain connectivity networks. Presented at The 4th Annual meeting of the COST Action "Analyzing the dynamics of information and knowledge landscape—KNOWeSCAPE", 22-24 February 2017, Sofia, Bulgaria.
23. Hassan, M., Dufor, O., Merlet, I., Berrou, C., and Wendling, F.: EEG Source Connectivity Analysis: From Dense Array Recordings to Brain Networks, PLOS one, 9(8), e0105041 (2014)

Gene expression in schizophrenia patients and non-schizophrenic individuals infected with *Toxoplasma gondii*

Aleksandra Uzelac[1], Tijana Štajner[1], Miloš Busarčević[1], Ana Munjiza[2], Milutin Kostić[2], Čedo Miljević[2], Dušica Lečić-Toševski[2], Nenad Mitić[3], Saša Malkov[3], and Olgica Djurković-Djaković[1]

[1] Center of Excellence for Food- and Vector-borne Zoonoses, Institute for Medical Research, University of Belgrade, Dr. Subotića 4, 11129 Belgrade, Serbia
aleksandra.uzelac@imi.bg.ac.rs
[2] Institute of Mental Health, School of Medicine, University of Belgrade, Palmotićeva 37, 11000 Belgrade, Serbia
[3] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
{nenad,smalkov}@matf.bg.ac.rs

Abstract. There is an increasing body of data suggesting the association of infection with the protozoan parasite *Toxoplasma gondii* and schizophrenia. In this study, we employed a combination of data mining and bioinformatics to investigate whether any genes from loci which harbor schizophrenia associated single nucleotide polymorphisms (SNP) are involved in the immune response to *Toxoplasma gondii* infection. Of the 208 unique protein coding genes in 22 schizophrenia associated loci, 108 differ in expression by at least 30% with respect to controls according to data mining of published microarrays. Functional annotation clustering of those genes was used to select HLA-DQA1, TAP1, TAP2, PSMB8, EGFL8, LY6G6C, C4A and CFB for expression validation by real time PCR in peripheral blood of schizophrenia patients and controls. Preliminary results suggest that the levels of expression of HLA-DQA1 and TAP2 differ in *T. gondii* infected schizophrenia patients from *T. gondii* infected individuals without schizophrenia.

Keywords: bioinformatics, data mining, gene expression, *Toxoplasma gondii*

1. Introduction

Through an ingenious combination of conclusions drawn from various studies, infection with the ubiquitous neurotropic parasite *T. gondii* has become associated with schizophrenia (SCZ). The working hypothesis is that *T. gondii* plays a role in the etiology of SCZ. The most cited evidence which gave rise to this hypothesis are repeated observations of neurological symptoms and unusual behavior in experimentally infected animals, accompanied by increased levels of dopamine [1–5]. The discovery of the parasites intrinsic ability to manipulate

levels of dopamine and GABA, which are key neurotransmitters dysregulated in SCZ, has incited research to determine whether this could be the mechanism by which it affects its hosts behavior [2, 6, 7]. Several studies have reported higher seroprevalence of anti *T. gondii* antibodies in SCZ patients [2–8]. Additionally, longitudinal clinical studies have shown that children from mothers with high specific anti *T. gondii* IgG titers have an increased risk of developing SCZ later in life [9]. Others have shown serointensity to be higher in SCZ patients infected with *T. gondii* when compared to infected individuals without SCZ [10]. Remarkably, pre-pulse inhibition of the startle reflex (PPI), which is deficient in SCZ patients [11–13], may also be deficient or dysregulated in people [14] and rodents [15] infected with *T. gondii*. Despite the fact that the etiology of SCZ is notoriously difficult to study due to the usual late onset, staggering diversity of clinical manifestations and complex genetics, some insights are emerging. GWAS studies have accumulated evidence which suggests that this disorder may be a result of distinct SNP patterns [16, 17]. A particularly strong association has been suggested for SNPs which occur in the MHC region, hinting at the possibility of immune involvement and thus again implicating infection as a potential contributor to the etiology [17]. The majority of SNPs discovered thus far tend to occur in non-coding regions, which complicates matters. This finding points to the conclusion that SCZ is unlikely to be caused predominantly by faulty or inactive proteins and that other factors such as alterations in gene expression, splicing and methylation could play a role. For the purposes of this study, we chose to look at the expression of selected genes which reside in SCZ associated loci with genome wide significance as determined by GWAS and are involved in the immune response to *T. gondii* infection in peripheral blood of infected and uninfected SCZ patients and non SCZ controls.

2. Materials and Methods

2.1. Data mining / Bioinformatics Approach

To obtain genes associated with SCZ, we looked at published GWAS data. One study was selected because it featured a multistage GWAS approach which resulted in 22 loci with genome wide significance for schizophrenia [17]. A number of the loci had been previously published, while 13 were novel. A list of genes residing in those loci was obtained using the UCSC genome browser. To investigate the expression of those genes in *T. gondii* infection, we mined published microarrays of murine brain homogenates (Br) and lymphocytes (Ly) during acute infection with *T. gondii* type II strain ME49 [18]. We selected a cutoff value of 30% change in expression to represent significant changes with respect to uninfected controls. The group of genes with expression levels above the cutoff in both Br and Ly was analyzed with the functional annotation clustering algorithm included in the DAVID 6.7 bioinformatics suite (https://david.ncifcrf.gov/). The following genes: HLADQA1, TAP1, TAP2, C4A, CFB, PSMB8, EGFL8, LY6G6C, from the most significantly enriched cluster were selected for expression analysis of human peripheral blood. 2.2. Gene expression analysis in a patient series The study group consisted of SCZ patients (n=101) recruited at the Institute

for mental health (IMH), Belgrade, during regular follow-up since Dec. 2013. All patients were tested for *T. gondii* infection during an exacerbation episode or when they presented for routine control (no first episode of SCZ). A series of age-matched patients screened for toxoplasmosis at the Serbian national reference laboratory for toxoplasmosis (NRLToxo) was included as a control group.

2.2. Toxoplasmosis diagnostic assays

Serology was performed using the VIDAS system (bioMerieux, France) for quantification of *T. gondii* specific IgG and IgM antibodies and the avidity of the *T. gondii* specific IgG antibody.

2.3. Real time PCR and relative quantification

Total RNA was isolated from the cellular fraction of a 5ml whole blood sample using TriZol reagent (Life Technologies, Grand Island, NY), and converted into first strand cDNA using the RevertAid kit (Life Technologies,Grand Island, NY). Real time PCR was performed with the HiGreen commercial mastermix (Life Technologies, Grand Island, NY) in a StepOnePlus instrument (Applied Biosystems, Foster City, CA, USA). Relative fold changes were calculated using the $\Delta\Delta$Ct method.

3. Results

The 22 schizophrenia loci we looked into contained over 480 annotated sequences, which included splice variants of a number of genes, pseudogenes and miRNAs. Despite a thorough search of the literature and online data repositories, we were unable to find a suitable human transcriptome dataset to mine for expression values of the sequences we obtained from the GWAS. Instead, we selected a murine transcriptome dataset which represents gene expression in Br homogenates and Ly of Balb/c mice in acute infection with the *T. gondii* strain Me49. In light of that, the sequences from the SCZ loci had to be converted to their murine homologs and as a result, all pseudogenes and miRNAs were eliminated and only major splice variants were retained. The number of sequences we could analyze was further reduced by the lack of appropriate oligonucleotide probes on the array. Ultimately, we were able to obtain expression levels for a total of 208 genes from SCZ associated loci. This number of genes however is still far too large to be assayed by real time PCR but also far too small to justify running microarrays. To further reduce the number, we applied a cutoff value of 30% change in expression with respect to uninfected controls. The cutoff value was selected after analyzing the fold change (FC) values of all 208 genes and represents a medium level of stringency. After the cutoff was applied, 108 genes remained, distributed as follows: 74 Br genes, 84 Ly genes and 45 BrLy genes. A schematic representation of the gene selection approach is given in Figure 1.

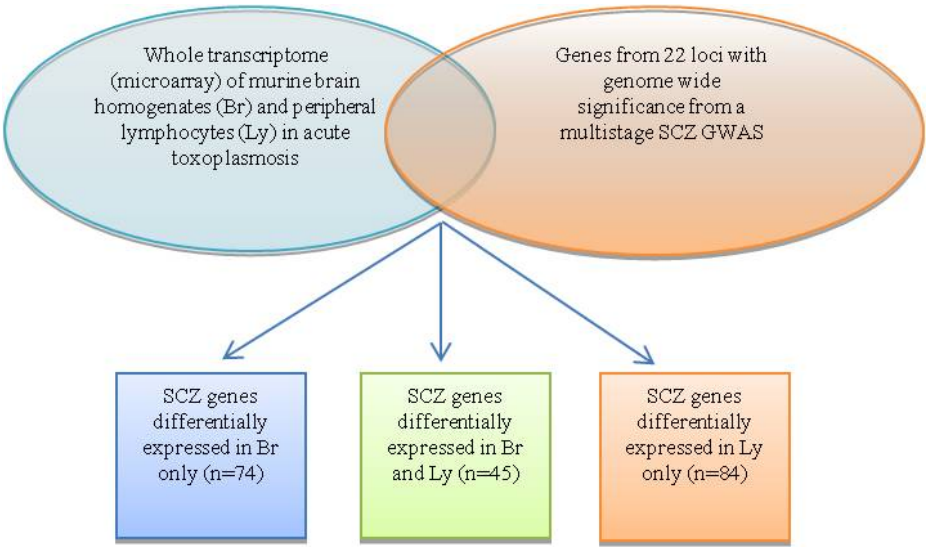


FIG. 1: Schematic representation of the gene selection approach

As our aim was to investigate gene expression of select genes by real time PCR in peripheral blood, we reasoned that the group of 45 genes, which is obtained by combining the Br and Ly datasets, contains the best candidates. In effect, by combining the two groups, we were able to eliminate most of the genes which are highly tissue specific and may not be expressed beyond baseline in peripheral blood. We next performed a functional annotation clustering analysis of this group of genes. The rationale for analyzing the genes by function was to gain insight into biochemical pathways or biological functions which are affected in SCZ and *T. gondii* infection. For that purpose, we selected the algorithm included in the DAVID 6.7 Bioinformatics suite. Functional annotation clustering results of the most significantly enriched cluster are shown in Figure 2.

The enrichment score we obtained for annotation cluster 1 is below the cutoff for statistical significance, most likely due to the very small number of genes included in the analysis. Nevertheless, the fact that the cluster is made up of genes with immune functions lends biological credibility to the results. In addition, we determined that most of the genes are situated in the MHC region, which has been repeatedly associated with SCZ by different GWAS, while its association with the immune response to infections cannot be disputed. The genes we ultimately selected to assay for expression in peripheral blood and the mined expression data are shown in Table 1.

Peripheral blood as well as demographic data was collected from a series of 101 SCZ patients of the Institute for mental health. All were outpatients whose blood was collected during one of the repeat check-up visits to the Institute. As

Annotation Cluster 1	Enrichment Score: 2.9473930058887032			
Term	Count	%	PValue	Fold Enrichment
GO:0006952~defense response	10	24.39	8.30E-07	8.46
GO:0051059~NF-kappaB binding	4	9.76	1.97E-05	71.63
GO:0006986~response to unfolded protein	4	9.76	2.93E-04	29.31
GO:0051789~response to protein stimulus	4	9.76	9.75E-04	19.45
GO:0007584~response to nutrient	4	9.76	0.002112	14.87
GO:0031667~response to nutrient levels	4	9.76	0.005524	10.56
GO:0009991~response to extracellular stimulus	4	9.76	0.007494	9.46
GO:0008134~transcription factor binding	5	12.20	0.013651	5.06
GO:0005829~cytosol	9	21.95	0.014536	2.62
GO:0010033~response to organic substance	5	12.20	0.041484	3.61

FIG. 2: Functional Annotation Clustering results, DAVID 6.7 Bioinformatics suite

TABLE 1: Selected genes with expression values and regulation in mined microarray datasets

Gene Symbol	FC (Br)	Regulation (Br)	FC (Ly)	Regulation (Ly)
HLA-DQA1	62.4	Up	3	Down
TAP1	24.3	Up	3.1	Up
TAP2	6.1	Up	2.2	Up
C4A	7.6	Up	4.9	Up
C4B	5.8	Up	23	Up
PSMB8	17.5	Up	2	Up
EGFL8	2.9	Down	1.4	Up
LY6G6C	5.2	Down	1.1	Down
CFB	156.6	Up	13	Up

a result, this study group did not include any patients with first episode SCZ. Specific anti *T. gondii* IgG was positive in 20 patients, which were thus classified as seropositive (toxo+). Negative specific IgM antibody as well as high avidity indices of the detected specific IgG antibodies indicated that 19 toxo+ patients were in the chronic phase of infection. Only one patient had serology findings which correspond to acute infection. Select demographic data for the seropositive (toxo+) and seronegative (toxo-) groups is shown in Table 2.

TABLE 2: Select demographic data of the schizophrenia patients enrolled in the study thus far

Charateristics	Seropositive (n=20)	Seronegative (n=81)
Age (yrs) Mean ± SD	48 ± 11.9	39.9 ± 9.4
Range	27-82	22-64
Sex M	14	46
F	6	35
Hereditary SCZ	4	21

Statistical analysis reavealed that the difference in the mean age between the toxo+ and toxo- groups is not significant. Importantly, the presence of SCZ in the family (hereditary SCZ) in the toxo+ group was 20% and 25.9% in the toxo- group. While the difference is not statistically significant, a higher prevalence of infection in the non-hereditary SCZ group may be interesting, and should be continued to be monitored.

Figure 3 shows the relative quantity of the HLA-DQA1, TAP2 and TAP1 genes determined by real time PCR in peripheral blood of the SCZ patients (SCZ+/toxo+ and SCZ+/toxo- groups) compared with a control group which consisted of individuals infected with *T. gondii* without SCZ (SCZ-/toxo+). For HLA-DQA1, there is a slight difference in expression between both SCZ+ groups and a clear difference with respect to the SCZ- control group. The same is evident for TAP2. However, there seems to be no difference in expression between any of the groups for TAP1. The expression of the remaining genes remains to be investigated.

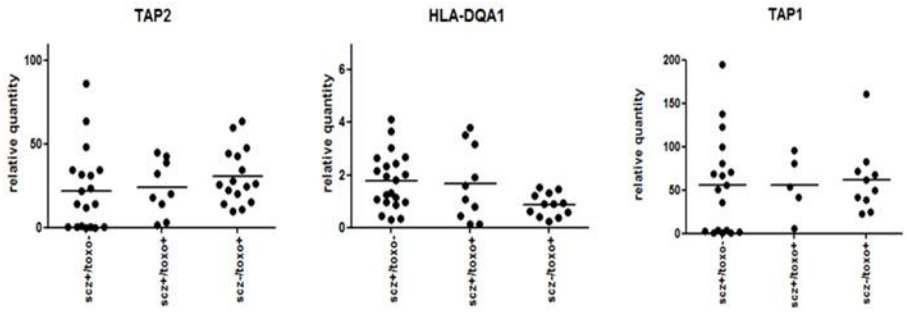


FIG. 3: Expression data for HLA-DQA1, TAP2 and TAP1 in peripheral blood

4. Discussion

This study highlights the use of a novel approach based on bioinformatics and data mining to investigate a complex and controversial field of research built

around the hypothesis that the parasite *T. gondii* plays a role in the etiology of SCZ. Research into SCZ etiology so far has yielded fairly convincing evidence for the existence of SNPs which occur statistically more frequently in the genomes of SCZ patients [17]. Given that most of these SNPs are not within coding sequences, it is reasonable to assume that their effect is exerted on gene expression and regulation [16]. We reasoned that by investigating gene expression in SCZ hosts who are infected with *T. gondii* and those who are not, we may gain insight into the mechanism(s) by which the parasite could be involved in the etiology of the disorder. Rather than employing methods which generate astronomical amounts of data which take years to process and analyze, we wanted a targeted approach which would allow us to analyze a small number of genes by real time PCR. The strategy we chose to apply for gene selection was utilizing and mining already published data. Instead of looking at whole transcriptomes which contain over 20,000 genes, we reasoned that by identifying genes in loci associated with SCZ by the presence of characteristic SNPs which are involved in the immune response to *T. gondii*, we should be able to narrow down the selection to those genes which highlight the impact of infection on the SCZ host. Combining GWAS data with microarray data may be an unusual approach, given the fundamental differences in the methodology and output data, but as the preliminary PCR results suggest, our results are biologically relevant. Functional annotation clustering performed as a control on the original list of coding sequences which we extracted from the 22 loci identified by the GWAS used, did not identify genes with immune functions as the top cluster. In fact, the results from this analysis yielded no biologically meaningful data, as may have been expected given that SCZ is such a diverse disorder in terms of genetics and clinical manifestations. It should not be surprising that the most significantly enriched cluster identified after combining GWAS and microarray data contained genes with immune functions, as most of the genes which are differentially expressed during infection are in fact those involved in the immune response. As a number of GWAS have identified a particularly strong association of SNPs distributed throughout the MHC locus with SCZ [19], the immune response may be a relevant factor in SCZ etiology and should not be ignored. Our preliminary expression data indicates that at least two of the genes we selected are indeed differentially expressed in the peripheral blood of SCZ patients with *T. gondii* infection compared to infected individuals without SCZ. It is particularly interesting to note that there also is a difference in the expression of these genes between SCZ patients with and without the infection, albeit far less striking. One tentative conclusion we can draw from the data thus far is that *T. gondii* infection changes the expression of genes which reside in SCZ associated loci as a consequence of the hosts immune response. Clearly, the infection alone is unlikely to cause the disorder, as the vastly different global prevalences of SCZ and toxoplasmosis indicateapproximately 1-2% compared to 30-50% of the global population, respectively [20, 21]. However, if the infection is introduced as a factor in a host already genetically predisposed to SCZ, it is conceivable that it could trigger the development of the disorder by inducing long term changes in gene expression and regulation. The longitudinal effect of *T. gondii* infection is particularly interesting given the fact that SCZ is predominantly a late (adult) onset disorder. Our

own data suggests that despite the fact that we had based our gene selection on microarray data from acutely infected animals, the genes which were differentially expressed in acute disease were still differentially expressed in our SCZ patients who were all in the chronic phase of infection.

The results presented here, albeit preliminary, have already shown biological significance. Final interpretation of the results and conclusions will be drawn in the future, when a greater number of SCZ patients are enrolled in this ongoing study.

References

1. Webster, J.P.: Rats, cats, people and parasites: the impact of latent toxoplasmosis on behaviour. Microbes Infect. 3:10371045 (2001)
2. Yolken, R.H., Dickerson, F.B., Torrey, E.F.: Toxoplasma and schizophrenia. Parasite Immunology 31:706-715 (2009)
3. Hermes, G., et al: Neurological and behavioral abnormalities, ventricular dilatation, altered cellular functions, inflammation, and neuronal injury in brains of mice due to common, persistent, parasitic infection. J. of Neuroinflamm. 5:48 (2008)
4. Halonen, S.K., Weiss, L.M.: Toxoplasmosis. Handb. Clin. Neurol. 114:125-145 (2013)
5. Brown, A.S.: The environment and susceptibility to schizophrenia. Prog. Neurobiol. 93(1):23-58 (2011)
6. Gaskell, E.A., Smith J.E., Pinney, J.W., Westhead D.R., McConkey, G.A.: A Unique Dual Activity Amino Acid Hydroxylase in Toxoplasma gondii. Plos One. 4(3):e4801 (2009)
7. Fuks, J.M., Arrighi, R.B.G., Weidner, J.M., Mendu, S.K., Jin, Z., Wallin R.P.A., Rethi, B., Birnir, B., Barragan, A.: GABAergig Signaling is Linked to a Hypermigratory Phenotype in Dendritic Cells Infected by Toxoplasma gondii. Plos Pathog. 8(12):e1003051 (2012)
8. Torrey, E.F., Bartko, J.J., Lun, Z.R., Yolken, R.H.: Antibodies to Toxoplasma gondii in Patients With Schizophrenia: A Meta-Analysis. Schizophrenia Bulletin 33(3):729-736 (2007)
9. Brown, A.S., Schaefer, C.A., Quesenberry, C.P.Jr., Liu, L., Babulas, V.P., Susser, E.S.: Maternal exposure to toxoplasmosis and risk of schizophrenia in adult offspring. Am. J. Psychiatry. 162:767-773 (2005)
10. Hinze-Selch, D., Daubener, W., Eggert, L., Erdag, S., Stoltenberg, R., Wilms, S.: A Controlled Prospective Study of Toxoplasma gondii Infection in Individuals With Schizophrenia: Beyond Seroprevalence. Schizophrenia Bulletin 33(3):782-788 (2007)
11. Braff, D.L., Stone, C., Callaway, E., Geyer, M.A., Glick, I., Bali, L.: Prestimulus effects on human startle reflex in normals and schizophrenics. Psychophysiol. 15:339 343 (1978)
12. Ludewig, K., Geyer, M.A., Vollenwender, F.X. Deficits in Prepulse Inibition and Habituation in Never-Medicated, First-Episode Schizophrenia. Biol. Psychiatry. 47:662-669 (2000)
13. Swerdlow, N.R., Weber, M., Qu, Y., Light, G.A., Braff, D.A.: Realistic expectations of prepulse inhibition in translational models for schizophrenia research. Psychopharmac. (Berl). 199(3):331-388 (2008)
14. Piplatov L., ebnkov, B., Flegr, J.: Contrasting Effect of Prepulse Signals on Performance of Toxoplasma-Infected and Toxoplasma-Free Subjects in an Acoustic Reaction Times Test. Plos One. 9(11): e112771 (2014)

15. Eells, J.B., Varela-Stokes, A., Guo-Ross, S.X., Kummari, E., Smith, H.M., Cox, E., Lindsay, D.S.: Chronic Toxoplasma gondii in Nurr1-Null Heterozygous Mice Exacerbates Elevated Open Field Activity. Plos One. 10(4): e0119280 (2015)
16. Harrison, P.J. Recent genetic findings in schizophrenia and their therapeutic relevance. J. Psychopharmacol. 29(2):85-96 (2014)
17. Ripke, S., et al: Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat. Genet. 45(10):1150-9. (2013)
18. Jia, B., Lu, H., Liu, Q., Yin, J., Jiang, N., Chen, Q.: Genome-wide comparative analysis revealed significant transcriptome changes in mice after Toxoplasma gondii infection. Parasit.Vectors. 4,6:161 (2013)
19. Bergen, S.E., Petryshen, T.L.: Genome-wide association studies (GWAS) of schizophrenia: does bigger lead to better results? Curr.Opin. Psychiatry. 25(2):76-82 (2012)
20. World Health Organization (WHO): www.who.int/mental_health/management/schizophrenia/en/. Accessed: April 28, 2017
21. Flegr, J., Prandota, J., Sovikov, M., Israili, Z.H.: Toxoplasmosis-A Global Correlation of Latent Toxoplasmosis with Specific Disease Burden in a Set of 88 Countries. Plos One. 9(3): e90203 (2014)

Viral: Real-world competing process simulations on multiplex networks

Petar Veličković, Andrej Ivašković, Stella Lau, and Miloš Stanojević

Computer Laboratory, University of Cambridge,
Cambridge CB3 0FD, UK
{pv273,ai294,sl715,ms2239}@cam.ac.uk

Abstract. Accurate modelling of spreading processes represents a crucial challenge of modern bioinformatics, particularly in the context of predicting the consequences of epidemics (e.g. the proportion of population infected at the critical point). A wide variety of frameworks have been established; especially, recent developments in *multiplex networks* allow for integrating several competing spreading processes and modelling their interactions more directly. However, the research developments so far have primarily been evaluated on randomly-generated networks and assumptions on network dynamics that are unlikely to correspond to actual human psychology. As a decisive step towards controlled experiments of this kind, we present *Viral*, a multiplex-network-guided system for real-world simulations of the competing processes of *epidemics* and *awareness* in modern society, based around a lightweight distributed Android application and a centralised simulation server, both of which are simple to set up and configure. Extensive logging facilities are provided for analysing the simulation results.

Keywords: multiplex networks, competing processes, epidemics, awareness, real-world simulations, Android

Availability: The full source code (licensed under the MIT license) is available at `http://github.com/PetarV-/viral`.

1. Introduction

Traditionally, epidemics modelling has been performed by way of *single-layered networks*, representing humans as nodes with a set of possible states they can be in (susceptible-infected-recovered (SIR) [1] and its varieties being a popular choice) and allowing for disease to spread along the links of the network, representing pairs of people that come into physical contact. However, it became apparent that one usually cannot fully describe the disease-spreading process in such an isolated fashion—there may well be other processes that are capable of *influencing* it (by way of either *competition* or *cooperation*), forming in that way a kind of *network of networks* [2]. One example of such influence can be an awareness-spreading process in the society, where people that are aware of an epidemic can take appropriate precautions (vaccination, protective masks, isolation, etc.), effectively reducing their probability of contracting the disease. Note also that awareness can spread along links that differ from the links used in the

original network—a simple example are social network acquaintances that do not interact physically. Representing the system in this way has demonstrated emergence of important phenomena that were not present in the single-layered case, therefore solidifying the need to focus on the "bigger picture" when performing disease modelling.

A popular interpretation of the concept of networks of networks is to represent them as *multilayer networks* [3]. In particular, the special case of *multiplex networks* should be very appropriate for representing epidemics-related networks and has been a topic of plentiful related research in recent years [4–10]. Informally, a multiplex network is a multi-layered graph in which each layer is built over the same set of nodes, and there may exist edges between nodes in different layers. Here the nodes usually represent individuals in a population, while the layers usually correspond to the different processes under study.

This framework has thus far been almost exclusively applied to generated networks (common choices include Erdős-Rényi random graphs [11] and Barabási-Albert scale-free networks [12]), and assumptions on the network dynamics (such as the Markov property) that may not always correspond to human psychology are often made. With the primary aim of providing a complementary tool that allows researchers to further verify their predictions on real-world controlled experiments, we have developed *Viral* during the 24-hour *Hack Cambridge* hackathon [13], where it was commended as one of the top seven projects (out of $\sim$100 participating teams from top-tier universities).

2. Multiplex network model

We consider a multiplex network setup with two layers over the same set of nodes (corresponding to individuals in the population), corresponding to the epidemics and awareness processes, respectively. An SIS (susceptible-infected-susceptible) process is assumed for the epidemics layer, while a UAU (unaware-aware-unaware) process is assumed for the awareness layer (akin to the models used in [4, 5]).

The epidemics layer operates under the assumption that the epidemic spreads by *airborne transmission*—each node broadcasts its current geolocation to the network, which then (re)computes mutual distances between nodes and normalises them to obtain probabilities of transmission (such that the disease is more likely to spread along pairs of nodes that are closer together).

Along the awareness layer, knowledge of an epidemic can spread between individuals that exchange information. Awareness is represented in this context by knowledge of a *round-unique code* (initially known to only a fraction of the population). We have made a crucial decision to model this layer only *implicitly*—individuals that have knowledge of this code may disseminate it by means of verbal communication as well as making advantage of social networks. This has a very positive impact on both the amount of state that needs to be maintained in the individual nodes and on obtaining a more accurate behavioural dynamic in this layer.

The layers influence one another in two critical ways: 1) a susceptible individual that knows the code can get *vaccinated*, thus diminishing their probability

of infection; 2) an individual that becomes infected will, with a fixed probability, obtain the round code, provided they didn't have it already. The full network dynamics are illustrated by Figure 1.

In order to discourage the "pack behaviour" in which awareness immediately fully spreads and everyone gets immunised early on, a novel component of our system encourages a proportion of the population to behave carelessly, by assigning them a negative role of an *infector*—their purpose being to get as much of the population infected as possible until the round ends. All other (*human*) nodes are simply tasked with staying healthy until the end of the round.

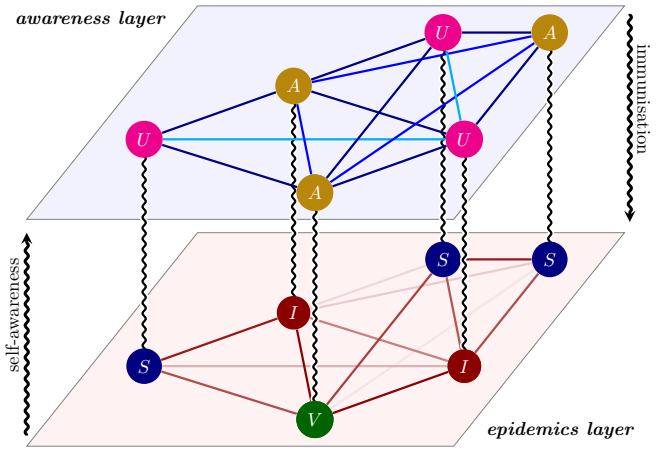


FIG. 1: Illustration of the underlying network dynamics assumed by *Viral*. Nodes take part in both the epidemics layer (SIS + vaccinated) and the awareness layer (UAU). Intensity of edges in the epidemics layer represents probability of contracting the disease over that edge, and is based on mutual proximity between the individuals. The dynamics of the awareness layer are more related to psychology, and therefore are modelled only implicitly to allow for maximal flexibility.

3. Implementation

Viral consists of two core components: the server and the Android application. The Android application represents a node in the network and broadcasts its current geolocation to the server, which is used to compute distances between nodes and obtain transmission probabilities for modelling the epidemics layer. The server simulates the state transitions (such as a change in awareness or physical state) and sends the node's updated state to the Android application.

3.1. Server

The server communicates with the Android applications as well as simulates and maintains the network state. It also periodically appends the network state into a log file for the current session and provides a visualisation tool that displays the most recent state (created in publication-ready Ti*k*Z format—examples can be seen in the synthetic experiments' outputs in Section 4.2).

Simulating the epidemics layer is achieved by maintaining a matrix $\mathbf{M}$ of inverse-exponential distances between all pairs of nodes with

$$\mathbf{M}_{ij} = ke^{-\lambda d_{ij}} \tag{1}$$

where $k > 0$ and λ are server parameters, and d_{ij} is the great circle distance between the locations of node i and node j, computed based on the "Inverse Formula" from [14]. The probability of activation for edge $i \leftrightarrow j$ is given by normalising:

$$\mathbf{P}_{ij} = \frac{\mathbf{M}_{ij}}{\sum_{i,j} \mathbf{M}_{ij}} \tag{2}$$

This means that the likelihood of infection increases as the *proximity* between nodes increases, corresponding to an assumption of *airborne transmission*. An edge activated between a susceptible and an infected node leads to the susceptible node becoming infected with a specified probability (also a server parameter).

The procedure as described requires $O(n^2)$ processing time complexity per each position update, as well as $O(n^2)$ space complexity (where n corresponds to the number of nodes in the system). While this should be sufficient for most applications, we have included the possibility of configuring the server to choose an edge to activate based on a *Gibbs' sampling* approach, reducing the storage requirement to $O(n)$, along with a greatly diminished time complexity per update (typically *near-constant-time*), making the system potentially scalable to *hundreds of thousands* of concurrent clients.

3.2. Android application

The Android application consists of two main graphical components (Figure 2):

- *Initial screen*: the first prompt which becomes visible to the user once the application is started; it allows the user to provide the *hostname* and *port* of a *Viral* server;
- *Main screen*: the screen responsible for showing all the necessary information received from the server, as well as allowing user input where necessary.

Once the the hostname and port are provided via the initial screen, all the necessary components of the application are initialised. Thereafter, messages from the server can trigger updates to the main screen. Concurrently, when the position of the device is changed, its new geolocation is submitted to the server. In addition, the user can enter (and potentially be shown) the round-unique code (as mentioned in Section 2) in order to initiate vaccination—this code can be shared among users, implicitly simulating the awareness layer.

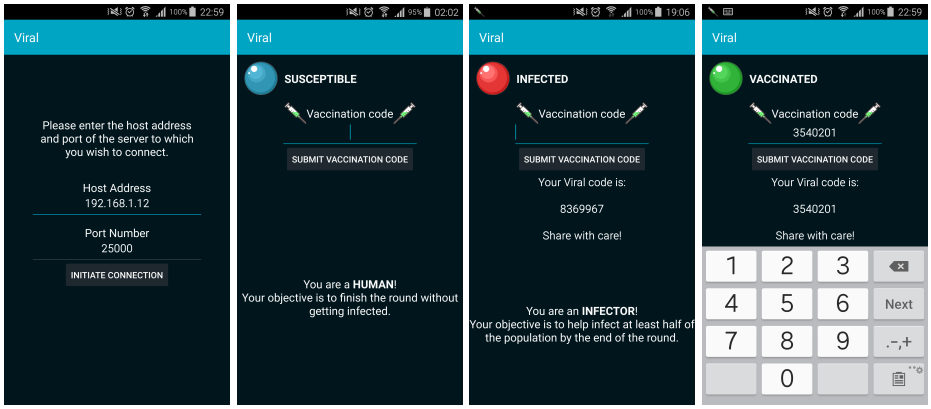


FIG. 2: A variety of screenshots of the *Viral* Android application. Left-to-right: the *initial screen*, followed by three different states of the *main screen*.

4. Usage

Viral is primarily pitched as a tool for performing *controlled experiments*; after setting up a central server (preferably on a UNIX-based system, to make advantage of the visualisation utility), participating subjects should have Android phones with the *Viral* application running in their possession, and understood the rules of the game. This section is primarily concerned with the necessary details for performing the initial setup and configuration of the server and the Android application. We also provide details of an auxiliary tool that can simulate additional participants performing a random walk and not doing any awareness-related interactions (other than vaccinating themselves if possible and desirable), and present a few results we have obtained in such synthetic experiments.

4.1. Installation

The full source code of *Viral* is hosted on the corresponding author's GitHub profile, at `https://github.com/PetarV-/viral`, and is licensed under the MIT license. The source may be downloaded as an archive from GitHub, or the repository may be directly cloned by running the following command within a terminal:

```
$ git clone https://github.com/PetarV-/viral.git
```

Detailed instructions for compiling and configuring the server, as well as setting up the Android application and configuring the synthetic clients used for the runs below, are provided in the README file of the repository.

4.2. Synthetic experiments

While the primary purpose of *Viral* is creating data from a controlled and real environment, it also supports the addition of bots (virtual participants), whom

the server does not distinguish from users. In the current model, the bots perform random walks and periodically send position updates to the server. Sending the updates is modelled as a Poisson process i.e. the time T between updates is a random variable with an exponential distribution $\mathscr{E}(\lambda)$, with the probability density function $f_T(t) = \lambda e^{-\lambda t}$. No other behaviour is given to the bots, other than them vaccinating themselves if they have access to the valid vaccine code and have the *human* role.

We have run our application on purely synthetic data for preliminary measurements. Some interesting cases of network behaviour (with different network parameters) can be seen in Figure 3.

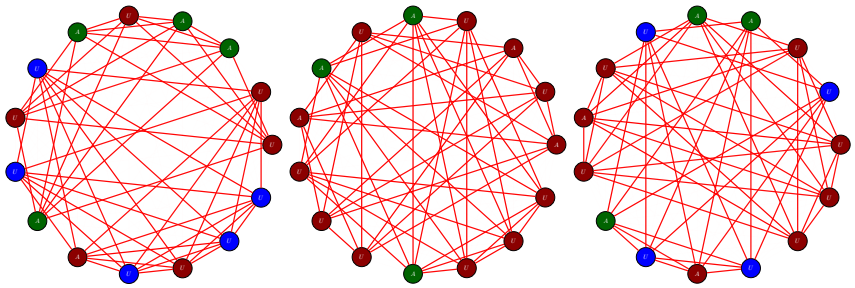


FIG. 3: Examples of round endings. The first diagram shows a situation with parameters corresponding to a typical flu-like epidemic (with an edge activation probability of 0.1 and a probability of contracting the disease of 0.05 if the individual is vaccinated). The second diagram corresponds to a pandemic-like scenario, in which everybody who is not vaccinated becomes infected (with an edge activation probability of 0.4). The third diagram corresponds to a severe epidemic with ineffective vaccines (with an edge activation probability of 0.25 and a probability of contracting the disease of 0.6 if the individual is vaccinated). The node colours correspond to the colour coding from Fig. 1, and the link intensities correspond to proximities in the epidemics layer. Note that in all these cases, the epidemics layer graph contains two high-proximity clusters.

5. Conclusions

In this applications note we have presented *Viral*, a utility for performing real-world controlled experiments on epidemics spreading with configurable parameters, taking advantage of the Android platform and multiplex networks. To the best of our knowledge, it is the first of its kind, and should serve as both a valuable tool for bioinformaticians and a potential reference implementation for future advancements in the area of real-world simultaneous spreading process simulation. In particular, the choice and amount of processes being considered should be extendable to other cases, such as simultaneously considering multiple transmission paths of a single disease [8] or multiple diseases [9]. We believe that the awareness component is also vital, and the framework provided

by *Viral* for implicitly simulating it should prove highly valuable in all future extensions. Furthermore, we hope that the *human/infector* model considered in Section 2 should be a valuable first step towards accurately simulating the fact that a large proportion of the population acts fairly carelessly in the presence of an epidemic.

6. Acknowledgements

We extend the deepest of thanks to the organisers of *Hack Cambridge 2016*, for giving us the initial opportunity and motivation to push this project forward in an extremely challenging setting, as well as to the event's judges for nominating us as one of the best projects at the event, and providing us with useful constructive criticism.

The work described within this manuscript has been preliminarily presented within the *Bioinformatics Seminar* at the Faculty of Mathematics of the University of Belgrade. We would hereby like to thank all of the attendees of the seminar session for their extremely useful comments, some of which have been since then integrated within *Viral*.

References

1. Kermack, W. O., & McKendrick, A. G. (1927, August). A contribution to the mathematical theory of epidemics. In Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences (Vol. 115, No. 772, pp. 700-721). The Royal Society.
2. Gao, J., Buldyrev, S. V., Stanley, H. E., & Havlin, S. (2012). Networks formed from interdependent networks. Nature physics, 8(1), 40-48.
3. Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., & Porter, M. A. (2014). Multilayer networks. Journal of Complex Networks, 2(3), 203-271.
4. Granell, C., Gómez, S., & Arenas, A. (2013). Dynamical interplay between awareness and epidemic spreading in multiplex networks. Physical review letters, 111(12), 128701.
5. Granell, C., Gómez, S., & Arenas, A. (2014). Competing spreading processes on multiplex networks: awareness and epidemics. Physical Review E, 90(1), 012808.
6. Buono, C., Alvarez-Zuzek, L. G., Macri, P. A., & Braunstein, L. A. (2014). Epidemics in partially overlapped multiplex networks. PloS one, 9(3), e92200.
7. Zhao, D., Wang, L., Li, S., Wang, Z., Wang, L., & Gao, B. (2014). Immunization of epidemics in multiplex networks. PloS one, 9(11), e112018.
8. Zhao, D., Li, L., Peng, H., Luo, Q., & Yang, Y. (2014). Multiple routes transmitted epidemics on multiplex networks. Physics Letters A, 378(10), 770-776.
9. Azimi-Tafreshi, N. (2015). Cooperative epidemics on multiplex networks. arXiv preprint arXiv:1511.03235.
10. Zuzek, L. G. A., Buono, C., & Braunstein, L. A. (2015). Epidemic spreading and immunization strategy in multiplex networks. In Journal of Physics: Conference Series (Vol. 640, No. 1, p. 012007). IOP Publishing.
11. Erdős, P., & Rényi, A. (1959). On random graphs. Publicationes Mathematicae Debrecen, 6, 290-297.
12. Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. science, 286(5439), 509-512.

13. Hack Cambridge. (n.d.). Retrieved March 29, 2016, from https://www.hackcambridge.com/
14. Vincenty, T. (1975). Direct and inverse solutions of geodesics on the ellipsoid with application of nested equations. Survey review, 23(176), 88-93.

DNA deformations as a tool for the genetic information implementation

Sergey N. Volkov

Bogolyubov Institute for Theoretical Physics, National Academy of Sciences of Ukraine,
Metrologichna Str. 12-b, 03680 Kiev, Ukraine
snvolkov@bitp.kiev.ua

Abstract. The accuracy of the genetic information transfer in living cells is largely due to the peculiarities of the structure and variability of the DNA double helix. A special role is played by deformations, which are induced by local conformational transformations in the macromolecule. The possible arising in this case the localized deformations provide wide tools in the processes of regulation and realization of the genetic information. The approach for the study of conformational depended deformations of DNA macromolecule is presented. This approach allows studying the mechanisms of the localized transformations of the double helix due to the action of small molecules, regulatory proteins, and some external forces on DNA structure. The obtained results give a consistent interpretation of the observed deformability of the regulatory fragments, such as TATA-box, A-tract, CpG steps, and provide the possibility to predict the sizes and energies of local deformation of the double helix at the location of some definite nucleotide sequences by them conformational states.

Keywords: Genetic information, DNA polymorphism, localized deformations

1. Introduction

Accuracy of the genetic information implementation in the living cells is largely due to the peculiarities of the structure and variability of the DNA double helix. The regulation of genetic activity, the stability and security of genetic texts, the reading and translation of genetic information, all of these important biological processes take place because of the unique properties of the DNA macromolecules, which distinguish them from other cellular molecules. The Nature use DNA molecules for the recording and storage of the genetic information, that is written by the sequences of DNA monomer units. To read and transfer the genetic texts the special informational system exists, which define the procedure and the need for the genetic information realization [1]. The key elements of these information system are the certain DNA fragments whose structure properties can be very different, and so can be used as structural signs for the correct understanding the genetic information.

As well known [2], the DNA double helix is a polymorphic macromolecule, and has the ability to change its secondary structure in dependence of the nucleotide composition or under an external action. The definite nucleotide se-

quences can take the unique deformations under interaction with proteins. Arising in this case the localized deformations provide a broad palette of tools in the genetic processes regulation.

The role of localized deformations caused by the polymorphic properties of the double helix is under thorough research lately [3, 4]. The thousands structures of DNA fragments interacting with proteins are collected in the Protei Data Bank and Nucleic Acid Database, but the mechanisms of the recognition of the definite DNA sites by the regulatory proteins remain not clear still. The point is that the studied molecular compounds have very a complicated structure, and the experimental methods by itself cannot give the understanding how these protein-DNA complexes form and work.

The theoretical methods have also some restrictions. The use of the computational approach based on quantum mechanical consideration of the interactions of all atoms in the macromolecule is restricted by the DNA fragments of a few base pairs [5]. In the case of using of atom-atom potential functions the computations may be fulfilled for greater DNA fragment, about 20 base pairs [6, 7]. But, the atom-atom potential parameters are sorted out for description of ground state of the system, and therefore these results cannot be considered as so correct for macromolecule metastable or excited configurations. The difficulties in the studding of the localized deformations are also connected to the fact that these deformations have sufficiently large amplitude of structural elements deviations from their equilibrium positions in the double helix, and therefore cannot be understood in terms of the elastic rod models, that are fair for the study of DNA chain mechanics in the harmonic approximation.

One of the productive way to understand the conformational resources of DNA double helix is the use of the phenomenological model approach. Under successful model construction the phenomenological approach allows to describe the structural transformations in macromolecules of the real size and with the account of surrounding factors. It is also important that the phenomenological approach make clear the physical nature of the processes under consideration. Of course, at some principal points, the phenomenological modeling can be amplified by more accurate calculations.

In the present work the phenomenological approach is used for the study of the conformationally induced deformations in the DNA regulatory fragments. Using the available results of the conformational analysis of the DNA regulatory sequences, their deformations is studied in the frame of two-component model. One model component (external) describes the macromolecule deformation as in the model of elastic rod, another component (internal) - the conformation transformations of the macromolecule fragment. Both components are considered as interconnected on the pathways of certain conformational transformation. The developed approach is used for a study of the deformability of TATA-box, A-tract, and the allosteric effects of CpG steps. The approach provides the possibility to predict the characteristics of local deformations of the double helix at the location of definite nucleotide sequences by them conformational states.

2. Conformation Transformations of the Regulatory Sequences in the Double-Stranded DNA

DNA macromolecule consists of two polynucleotide strands which are formed by alternated phosphates groups and deoxyribose (or sugar) rings. To the sugars the nucleic bases are attached. Under natural conditions, in ion-hydrate environment, DNA strands comprise a double helix, where the nucleic bases of different strands form the hydrogen-bonded complementary pairs (A·T or G·C) inside the helix. So, the double stranded DNA is a polymeric molecule, which monomer link includes two phosphate groups, two sugar rings and complementary pair of nucleic bases [2]. It is important, that DNA macromolecule has the ability to change the double helix form under the influence of external factors. Under the form transformation the macromolecule is usually change the conformation of the monomer link, that is, exhibits polymorphic properties. As in any polymorphic structures, the form of DNA double helix is coupled with the conformational state of its monomer link [2].

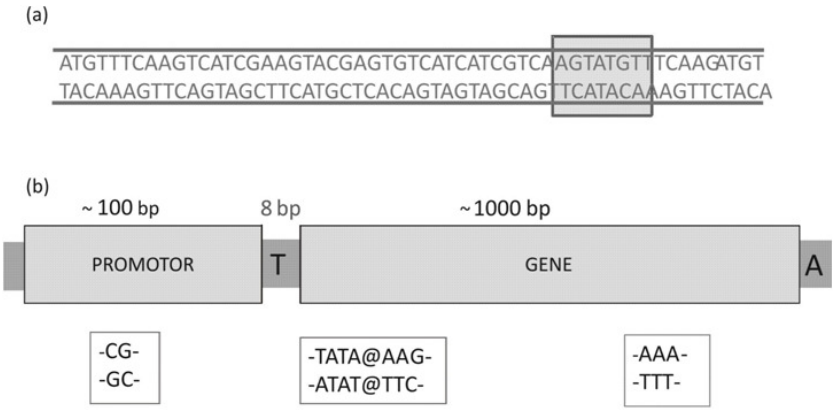


FIG. 1: Nucleotide sequences for DNA genetic activity regulation. (a)The regulatory sequences in the double-stranded DNA include a few base pairs; (b)CpG steps are situated usually at the promotor region; TATA-box determine the start of the transcription; A-tract points to the end of the gene.

In the present work we will pay attention to the deformations of the definite nucleotide sequences, which are usually used for the regulation of the genetic processes providing in DNA. There are the sequences of the TATA-box, A-tract, CpG-steps. These sequences are located in the gene promotor, or at the start and at the end of the nucleotide sequence of the gene (Fig. 1). The regulatory fragments in the DNA chain determine the order in which the genetic information should be read.

General property for these regulatory sequences is the polymorphism of their secondary structures in the double helix [8–10]. That is, the DNA fragments, where such sequences are located, have a several stable conformational states,

usually one ground state and some metastable states. Under the interaction with proteins or small ligands the regulatory sequences can transmit in another conformational state and induce the deformation caused by that conformational rearrangements. As a rule, the deformations of the regulatory fragments have a unique form and play a role of the special signals for the genetic information transfer.

They demonstrate the unique deformations of the double-stranded chain, which allows them to be used as regulatory elements in the mechanisms of genetic activity.

It is known, the position of TATA-box determines the beginning of the gene sequence and the start of the transcription [8, 10]. The results X-ray studies [8, 11, 12] are shown the transformation of the TATA-box under the action of RNAp. After the interaction with the polymerase subunits the TATA-box bends significantly (about 80° on eight base pairs), and by this means indicates the place of the transcription beginning. From the point of view of A.Klug [8] the anomaly deformability of the TATA-box is caused by a specific heteronomic conformation of the TATA sequence, where the conformations of the monomer units alternate. The X-ray study of the TATA-box transformations has confirmed this assumption and shown the existence of alternating in the conformations of the sugar rings (C'2-endo - C'3-endo) in the double helix strands [12].

In turn, as is well known [13], the fragment with A-tract is usually bent at the physiological conditions, and in such a way determine the finish point of the transcription. The bent state of A-tract is caused by the formation of water spine in the DNA minor groove with A·T pair sequences [14]. As shown by [14], after the water molecules separation the sequences of A-tract transit to *B*-like form. That is the bend of the tract is its usual state in the physiological conditions, and *B*-form is metastable for it.

The possible role of CpG-steps is the blocking of the genetic processes after cytosine methylation, and it is manifested mainly in promotor region [15]. The study of the flexibility of CpG-steps under the methylation of cytosine base shows the increase in rigidity of the DNA helix, and is seen directly in MD simulations [16]. It should be noted, that the methylated state of the step is less profitable by the energy. The experimental studies have shown that the CpG steps in the promotor region are located at some distance from each other [15]. That is, between two neighboring CpG-steps in the DNA chain there is an action of allosteric type on some advantageous distance between them.

Despite a significant number of studies performed, the origin of these unusual deformations has not been clarified yet. It should be emphasized, the study of the conformations of these fragments has shown their evident bimodality in the conditions of their functioning. So, it is correct to assume that the unusual features of these sequences are associated with their conformational polymorphism. In each case one can see the existing of the conformational transition from ground to some metastable state, that determine the appearance of unique deformation of the DNA fragment. It is important to note, that the conformational induced deformations are formed in the DNA fragments, where the definite nucleotide sequences are placed. Hence, these deformations have a localized character.

3. Two-component Approach

For understanding the mechanism of the localized deformations appearance and for the determining the shape of deformed DNA fragments the two-component approach will be used. The approach is developed in our earlier works for the study the conformational induced deformations in DNA macromolecules [17–19]. In this approach the transformation of DNA structure is considered in the frame of the model with two displacement component. One model component (external) describes the macromolecule deformation as in the model of elastic rod, another component (internal) - the conformation changes of the macromolecule monomer units. In fact, the model components describes the displacements of the monomer unit in the double helix and the displacement of the nucleic base pairs with respect to the DNA backbone. Both components are treated as interconnected on the paths of certain conformational transformation.

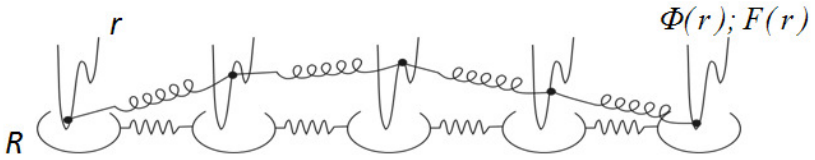


FIG. 2: Two-component model for the description of the deformations of a polymorphic DNA. R and r - the external and internal displacements of the macromolecule monomer unit. Φ and F - the potential functions describing the conformational transition between the ground and metastable states.

Taking into account the structural organization of a macromolecule, the energy density of the polymorphic macromolecule can be presented in the continuum approximation as:

$$\mathscr{E}_{def} = S_R R'^2 + s_r r'^2 + \Phi(r) - \chi h F(r) R' \quad . \tag{1}$$

In the expression (1) $R = R(z)$ and $r = r(z)$ are the displacements of the external and the internal components, R' and r' - its derivations on z, S_R and s_r - the elastic parameters of the external and the internal subsystems. The potential functions $\Phi(r)$ and $F(r)$ describe the conformational transition between the different conformational states of the macromolecule, the parameter χ determines the interrelations between the external and internal components of the macromolecule deformation. The parameter of the interrelations can have the different sign, and so it will be assumed that $\chi = i\chi_o$, where $i = \pm 1$ in dependence of the known form of the potential energy of the deformation. The value h is the size of the double helix step - the distance between the neighboring base pairs.

The first term in the equation (1) describes the energy of the macromolecule deformation, as in the elastic rod model. The second term describes the elastic energy of the rearrangements of the internal structure of the double helix. The

third term is the energy of the conformational transformation in the double helix structure. The last one describes the energy of the interrelation between the internal conformational rearrangements and the deformations of the double helix as a whole. Writing the energy of the deformation of the macromolecule in the form (1) supposes that the internal conformational energy has a double-well form, describing the conformational transitions from ground (usually r_0) to metastable (r_2) state and back. It is reasonably to consider that these two states are separated by the potential barrier placed between them (r_1). For more general character of the consideration we will study the possible deformations of the DNA fragments without definition of explicit expressions for the potential functions. Under study of the deformation of the definite DNA fragment it will be stipulated the type of the potential functions $\Phi(r)$ and $F(r)$ and the possible pathway of the deformation. For the ground state $r = r_0$ (DNA B-form) we will assume that:

$$\Phi(r_0) = F(r_0) = 0. \tag{2}$$

Let us find the static excitations of the polymorphic DNA fragments. The equations for the static excitation of the macromolecule with the energy (1) for the external and the internal components have the form:

$$R'' + \chi_1 \frac{dF}{dr} r' = 0\,; \tag{3}$$

$$r'' - \frac{\sigma_r}{2} \frac{d\Phi}{dr} - \chi_2 \frac{dF}{dr} R' = 0\,. \tag{4}$$

where $\sigma_r = 1/k_r h^2$, $\chi_1 = i\chi/hk_1$ and $\chi_2 = i\chi/hk_2$. After one time integration of the equation (3) with accounting of the condition (2) one can obtain the expression for the external deformation of the macromolecule:

$$\rho(z) = R'h = \frac{i\chi}{k_R} F(r)\,. \tag{5}$$

As is seen, the deformation of the macromolecule is determined by the form of the solution for the conformational component. That solution can be found from the equation, which is the result of the integration of the expression (4):

$$r'^2 + Q(r) = 0\,. \tag{6}$$

In the expession (6):

$$Q(r) = -\sigma_r \big[\Phi(r) - \frac{2\chi^2}{k_R} F^2(r)\big] - C\,. \tag{7}$$

Here C - the constant of integration, which is determined by the boundary conditions of the solution.

The equation (6) gives the form of the internal component of the considered fragment, its conformational state. This solution can be found as a conversion of the integral:

$$\int_{r(0)}^{r(z)} \frac{du}{\sqrt{-Q(u)}}\,. \tag{8}$$

The obtained expressions determine the form of the deformation of the macromolecule fragment and the conformation state of its internal structure, and can be used for the study of the deformability of the DNA regulatory sequences.

The deformation of the macromolecule chain in the extrema points will be determined by the following expressions:

$$\rho(z) = \frac{i\chi_o}{k_R} F(r) \,, \tag{9}$$

where i=$\pm$1, as is stipulated in the previous section.

4. Static Excitations of the DNA Regulatory Fragments

Let us find the possible static excitations for the macromolecule fragments with different shapes of the conformational energy, that is determined by the form of the function $Q(r)$.

As was mentioned above, the regulatory sequences in DNA macromolecule have a polymorphic properties. Thus, in modeling the conformational transformations of these sequences we will assume that their conformational energies have the definite bimodal shapes. This assumption is completely consistent with the experiment and computational modeling data [8, 11, 12, 14, 16]. In the study of the conformationally induced deformation of the DNA fragments it is reasonable to consider three cases, in accordance with the possible types of the bimodality which can realized in bistable systems. The different shapes of the conformational energy of the two-component fragments is presented in Fig. 3.

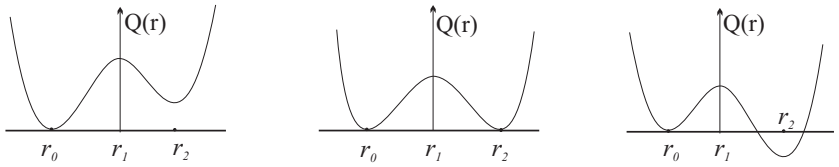


FIG. 3: The shapes of the conformational energy of the regulatory fragments in the double-stranded DNA.

So, for the modeling the deformation induced by the CpG steps the double-well potential with ground and metastable states should be used (Fig. 3a). The ground state will describe the usual conformation of the step in the double helix (*B*-form) and the metastable state let describe the methylated conformation.

For the study the deformation induced by the TATA-box the double-well potential should be presented in the form with two equivalent by energy states (Fig. 3b). Accordingly to our analysis of the experimental data in [19], in the TATA-box two equal by the energy conformations can be realized.

In the modeling of A-tract transformations the double-well potential with ground and metastable states should be used also. But in this case it should be

taken into account that metastable state in A-tract is more profitable due to the existence of the water spine in the DNA grove (Fig. 3c).

In accordance with the described properties of the DNA regulatory fragments and the possible types of their bimodality (Fig. 3), let us consider the conformationally induced deformations for three different conditions:

First, when the ground state accords to: $r = r_0;\ \Phi(r_0) = F(r_0) = 0$, and the metastable one is: $r = r_2;\ \Phi(r_2) > 0,\ F(r_2) > 0$.

Second, when the ground state accords to: $r = r_0;\ \Phi(r_0) = F(r_0) = 0$, and the metastable state is:$r = r_2;\ \Phi(r_2) = F(r_2) = 0$.

Third, when the ground state accords to: $r = r_0;\ \Phi(r_0) = F(r_0) = 0$, and the metastable one is: $r = r_2;\ \Phi(r_2) < 0,\ F(r_2) < 0$.

It is important for the understanding of the value and the shape of the DNA fragment deformations in accordance with the formula (9) to take into account the shape of the conformational energy of the fragment. For the first two cases correctly to assume: i=+1, but for the last case it is i=-1 in accordance with conformational analysis [14].

For the **first** case the equations for the internal components ($r(z)$) gives the following solution:

$$r(z) = r_2 - r_{ex}(z);\ r_{ex}(z) = \frac{a}{\mathbf{ch}(q_b z) + b}\ , \tag{10}$$

accordingly to which when $z \to 0$, then $r(z) \to r_2 - e_2$, and $\rho_{ex} \to \rho_e << \rho_2$. But when $z \to \pm\infty$, then $r(z) \to r_2$, and $\rho_{ex} \to \rho_2$. This solution can be used for the interpretation of the deformation of the DNA fragment with CpG steps, which are located on the edges of the fragment.

For the **second** case, one can obtain:

$$r(z) = r_1 + e\,\mathbf{th}(q_e z);\ , e = (r_2 - r_0)/2,\ q_e = \pm e Q_e\ , \tag{11}$$

which shows that under $z \to 0$, $r(z) \to r_1$, and $\rho_{ex} \to \rho_1$. In turn, when $z \to \pm\infty$, then $r(z) \to r_0$, and $\rho_{ex} \to 0$. This case accords to the understanding of the deformation of the TATA-box.

For the **third** case, one can obtain:

$$r(z) = r_1 + r_{ex}(z);\ r_{ex}(z) = \frac{c}{\mathbf{ch}(q_d z) + d}. \tag{12}$$

As is seen, in this case when $z \to 0$, then $r(z) \to r_0 + e_0$, and $\rho_{ex} \to \rho_0 \sim \rho_2$. But, when $z \to \pm\infty$, then $r(z) \to r_0$, and $\rho_{ex} \to 0$. This solution allows to understand of the deformation induced by A-tract.

All these solutions for the definite shapes of the conformational state and induced deformations are presented in Fig. 4.

As is seen from the presented in Fig. 4 results, the localized deformation of the definite fragments of DNA double strand are the static conformational excitations - static conformational solitons by their nature. The appearance of these excitations under definite boundary conditions serve as a specific signal for the DNA recognized proteins.

Thus, the CpG steps after methylation become more rigid and their placement in the promoter region leads to an increase in the stiffness of the nucleotide

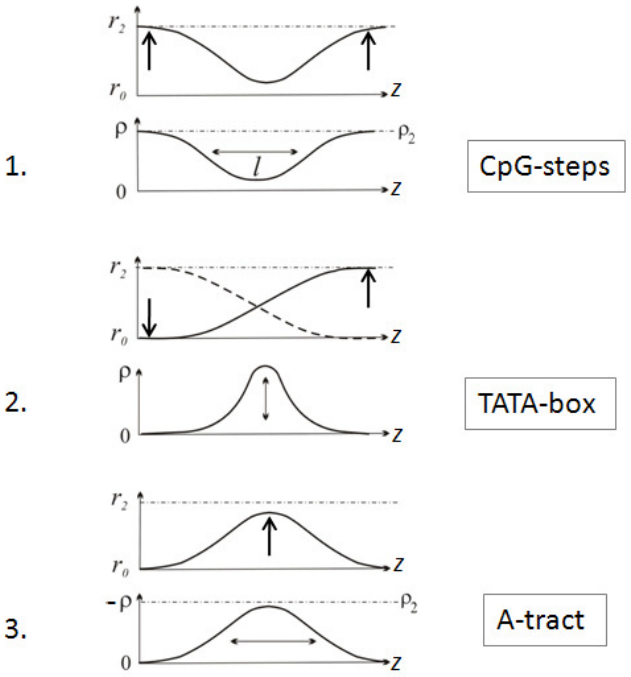


FIG. 4: Deformation state of DNA regulatory fragments for three studied shapes of the conformational energy. The displacements of the internal (conformation - $r(z)$) and external (deformation - ρ) components are shown.

sequences of the promoter itself and the impossibility of promotor recognition by the regulatory protein. This effect is enhanced by the location of methylated CpG steps at a certain distance (shown in Fig. 4.1 by arrows), the value of which corresponds to the width of the static soliton (or to its two half-widths), parameter l in Fig. 4.1. The boundary conditions on the ends of the fragment accord to the methylated CpD step itself, and in the central part of the fragment the value of the deformation drops to its usual value. The location of the two localized excitations (induced by two methylated CpG steps) on the favorable distance leads to the formation of additional tension in the promotor region and to the blocking of the genetics activity.

On the other hand the bistability of the TATA-box leads to the formation of a special deformation of the chain of the macromolecule with the maximum value in the centre of fragment (Fig. 4.2), Such shape of the fragment deformation allows to recognize this sequence with an accuracy of one base pair in DNA. It is necessary to emphasize that in the case of the TATA-box the special conditions on the ends of the fragment it is necessary to create. As is seen from the our theory, it is necessary to form different conformational state at the ends of the fragment (as shown in Fig. 4.2 by arrows). These boundary conditions are satisfied to real situation in TATA-box, due to the transitions of the terminal sugar rings of the

DNA fragment to different conformational states, as is shown in our analysis [19].

The conformation of A-tract is due to the interaction with water molecules on the length of the tract. This additional interaction leads to a decrease in the energy of the fragment accordingly to [14]. In the same time on the free ends of the tract the usual B-form should be realized, and that is the boundary conditions for the excitations of A-tract (Fig. 4.3). As a result, the A-tract is bent sufficiently smoothly in accordance with the number of A·T pairs interacting with water molecules (Fig. 4c).

Thus, the developed approach allows studying the physical mechanisms of the localized deformations appearance in the DNA double helix and shows the relationship of deformation with the conformational state of the DNA fragment and its nucleotide composition. As is seen, the mechanisms of the localized deformation conditioned by the nonlinear ability of the conformational transformations in the DNA regulatory fragments. The key role in the existence of this mechanisms play the polymorphic properties of the double helix. It is important to emphasize that polymorphism of the secondary structure is inherent in the DNA molecule, which apparently is the reason for its central position in the molecular biology.

5. Acknowledgement

The author is grateful to the Organizing Committee of the Belgrad BioInformatics Conference for the invitation to participate in its work and hospitality.

References

1. Ptashne, M.: A Genetic Switch. Phage Lambda Revisited. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA. (2004)
2. Saenger, W.: Principles of Nucleic Acid Structure. Springer-Verlag, New York, USA. (1984)
3. RCSB Protein Data Bank: www.rcsb.org
4. Nucleic Acid Database: http://ndbserver.rutgers.edu
5. Sponer, J., Hobsa, P.: Structure, Energetics, and Dynamics of the Nucleic Acid Base Pairs: Nonempirical *Ab Initio* Calculations. Chem. Rev., Vol. 99, 3247-3276. (1999)
6. Lavery, R.: Modeling Nucleic Acids: Fine Structure, Flexibility, and Conformational Transitions. Adv. Comput. Biol., Vol. 1, 69-145. (1994)
7. Auffinger, P., Westhof, E.: Simulations of the Molecular Dynamics of Nucleic Acids. Current Opinion in Structural Biology, Vol. 8, 227-232. (1998).
8. Klug, A.: Opening the Gateway. Nature (London), Vol. 365, 486-487. (1993)
9. Dickerson, R.E., Chiu, I.K.: Helix Bending as a Factor in Protein/DNA Recognition. Biopolymers, Vol. 44, 361-374. (1997)
10. Nikolov, D.B., Burley, S.K.: RNA Polymerase II Transcription Initiation: A Structural View. Proccedings of the National Academy of Sciences of USA, Vol. 94, 15-26. (1997).
11. Kim, Y., Geiger, J.H., Hahn, St., Sigler, P.B.: Crystal Structure of a Yeast TBP/TATA-box Complex. Nature, Vol. 365, 512-520. (1993)
12. Kim, J.L., Nikolov, D.B., Burley, St.K.: Co-crystal Structure of TBP Recognizing the Minor Groove of a TATA Element. Nature, Vol. 365, 520-527. (1993)

13. Segal, E., Widom, J.: Poly(dA:dT) Tracts: Major Determinants of Nucleosome Organization. Current Opinion in Structural Biology, Vol. 19, 6571. (2009)
14. Beveridge, D.L., Dixit, S.B., Barreiro, G., Thayer, K.M.: Molecular Dynamics Simulations of DNA Curvature and Flexibility: Helix Phasing and Premelting. Biopolymers, Vol. 73, 380403. (2004)
15. Lovary, P.T., Widom, J.: New DNA Sequences Rules for High Affinity Binding to Histone Octamer and Sequence-directed Nucleosome Positioning. Journal of Molecular Biology, Vol. 276, 19-42. (1998)
16. Perez, A., Castellazzi, Ch.L., ... and Orozco, M.: Impact of Methylation on the Physical Properties of DNA. Biophysical Journal, Vol. 102, 21402148. (2012)
17. Volkov, S.N.:. Modeling Large-scale Structure Dynamics of DNA Macromolecules. Biophysical Bulletin (Kharkiv, Vol. 7, 7-15. (2000); Modeling DNA Structural Transformations. ibid, Vol. 12, 5-12. (2003); arXiv/q-bio/BM0312034. (2004)
18. Volkov, S.N.: Modeling *B-A* Transformation of the DNA Double Helix. Journal of Biological Physics, Vol. 31, 323-337. (2005)
19. Kanevska, P.P., Volkov, S.N.: Intrinsically Induced Deformation of a DNA Macromolecule. Ukrainian Journal of Physics, Vol. 51, 1003-1007. (2006)

Author Index

Andjelković, Miroslav, 134
Aru, G. F., 18
Avetisov, Vladik, 1

Banović, Bojana, 102
Barcaro, Umberto, 108
Beljanski, Miloš, 73, 102
Borshcheva, Ekaterina, 1
Bugay, A. N., 18
Busarčević, Miloš, 142

Carboncini, Maria Chiara, 108

Djurić, Tamara, 82
Djurković-Djaković, Olgica, 142
Dragovich, Branko, 29
Dudić, Dragana, 102
Dushanov, E. B., 18

Filipović, Vladimir, 43
Friedmann, Naama, 88

Graovac, Jelena, 38
Grbić, Milana, 43

Ivašković, Andrej, 151

Jandrlić, Davorka, 55
Jelić, Asja, 64
Jelović, Ana, 73
Jovanović, Ivan, 82
Jovanović, Jasmina, 82

Kartelj, Aleksandar, 43
Kostić, Milutin, 142
Kovačević, Jovana, 38

Lakretz, Yair, 88
Lau, Stella, 151
Lečić-Toševski, Dušica, 142

Malkov, Saša, 142
Matić, Dragan, 43
Mišić, Nataša, 29
Miljević, Čedo, 142
Mitić, Nenad, 55, 73, 142
Munjiza, Ana, 142

Nicolaidis, Argyris, 93

Pajić, Vesna, 102
Paradisi, Paolo, 108
Parkhomenko, A. Yu., 18
Pavlović Lažetić, Gordana, 38
Pavlović, Mirjana, 55

Righi, Marco, 108

Salvetti, Ovidio, 108
Sciarrino, A., 119
Sebastiani, Laura, 108
Sorba, Paul, 119
Stanković, Aleksandray , 82
Stanojević,Miloš, 151

Štajner, Tijana, 142

Tadić, Bosiljka, 134
Treves, Alessandro, 88

Uzelac, Aleksandra, 142

Veličković, Petar, 151
Virgillito, Alessandra, 108
Volkov, Sergey, 159

Živković, Maja, 82

S P O N Z O R S


CEI
CENTRAL EUROPEAN INITIATIVE

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

Telekom Srbija

Telekom Srbija


SevenBridges


RNIDS
Register of National Internet
Domain Names of Serbia


GenomiX
4Life
MICROARRAYS
SEQUENCING
BIOMARKERS
BIONFORMATICS

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

57+61]:004(082)

BELGRADE BioInformatics Conference (2016 ; Beograd)
Proceedings / Belgrade BioInformatics Conference 2016,
20 - 24 June 2016, Belgrade, Serbia ;
[editor Nenad Mitić]. - Belgrade :
Faculty of Mathematics, University, 2017
(Belgrade : Donat Graf). - [12], 168 str. ; 25 cm

Str. [9-10]: Preface / Branko Dragovich, Gordana Pavlovic-Lazetić, Nenad Mitić. -
Tiraž 100. - Napomene i bibliografske reference uz radove. -
Bibliografija uz svaki rad. - Registar.

ISBN 978-86-7589-124-6

a) Биомедицина - Информациона технологија - Зборници

COBISS.SR-ID 254272012