Book of Abstracts

Conference
Data Mining in Bioinformatics

26-28 June 2012, Belgrade, Serbia

UNIVERSITY OF BELGRADE
FACULTY OF MATHEMATICS

Nenad Mitić, editor

Conference Data Mining in Bioinformatics 2012

Book of abstracts

Belgrade, June 26th-28th

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 University of Belgrade - Institute for General and Physical Chemistry, Faculty of Biology, Faculty of Physical Chemistry and Institute of Physics.

Sponsoring Institutions

The conference is financially supported by

- Ministry of Education and Science of the Republic of Serbia
- Postal Saving Bank J.S.C., Serbia
- P. E. of PTT Communications "Srbija"
- RNIDS - Register of National Internet Domain Names of Serbia

Publication of this Book of abstracts is financed by the Ministry of Education and Science of the Republic of Serbia

Publisher: Faculty of Mathematics, University of Belgrade

Printed in Serbia, by University of Belgrade, - Faculty of Technology and Metallurgy, Belgrade

Serbian National Library Cataloguing in Publication Data
Faculty of Mathematics, Belgrade
Book of Abstracts: Data Mining in Bioinformatics, 26-28 June 2012.– Book of abstracts
Nenad Mitić, editor. XIV+53 pages, 24cm.

ISBN: 978-86-7589-085-0

Number of copies printed: 100

Program Commitee

Gordana Pavlović-Lazetić	Faculty of Mathematics, University of Belgrade, Serbia
Zoran Obradović	Centre of Data Analytic and Biomedical Informatics, College of Science and Technology, Temple University, USA
Peter Tompa	VIB Department of Structural Biology, Brussels, Belgium and Institute of Enzymology, Hungarain Academy of Sciences, Budapest, Hungary
Vladimir Brusić	Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston MA and Health Informatics Group, MET Computer Science, Boston University
Nataša Pržulj	Department of Computing, Imperial College London, UK
Miloš Beljanski	Institute of General and Physical Chemistry, University of Belgrade, Serbia
Nenad Mitić	Faculty of Mathematics, University of Belgrade, Serbia
Mirjana Pavlović	Institute of General and Physical Chemistry, University of Belgrade, Serbia
Saša Malkov	Faculty of Mathematics, University of Belgrade, Serbia

Organizing Commitee

Gordana Pavlović-Lazetić	Faculty of Mathematics, University of Belgrade, Serbia
Miloš Beljanski	Institute of General and Physical Chemistry, University of Belgrade, Serbia
Ljiljana Kolar-Anić	Faculty of Physical Chemistry, University of Belgrade, Serbia
Branko Dragović	Insitute of Physics, University of Belgrade, Serbia
Nenad Mitić	Faculty of Mathematics, University of Belgrade, Serbia
Mirjana Pavlović	Institute of General and Physical Chemistry, University of Belgrade, Serbia
Saša Malkov	Faculty of Mathematics, University of Belgrade, Serbia
Jovana Kovačević	Faculty of Mathematics, University of Belgrade, Serbia
Biljana Stojanović	Faculty of Mathematics, University of Belgrade, Serbia

Preface

The Bioinformatics Research Group of the Faculty of Mathematics celebrates its tenth anniversary by organizing the first International Meeting on Data Mining in Bioinformatics. The purpose of this meeting is to bring together researchers interested in the development and application of data mining and information technologies to the field of life sciences. The focus of this meeting is on data mining. However, areas of interest include other information technologies, as well, such as sequence analysis, biostatistics, pattern recognition, machine learning and other related fields. Specific purpose of the meeting is to promote bioinformatics among researches and prospective researches here in Belgrade and Serbia, broadening and deepening our own understanding of the field and bringing together different research communities, from informatics to molecular biology to biomedicine, thus helping us consider possibilities for future cooperation with colleagues from different fields.

The meeting is of a workshop kind, with invited lectures only. We are grateful to all the participants who accepted our invitation to present their research. The book of abstracts of their presentations is in our hands. We thank to the Ministry of Education and Science of the Republic of Serbia for financially supporting publication of this book of abstracts. We also thank all who helped us in making this event happen.

We shall do our best to make the Data Mining in Bioinformatics conference become a tradition and a place of gathering, exchange and birth of ideas in the field of bioinformatics here in Belgrade, as well as a place where we all feel good, as being at home.

June, 2012

Gordana Pavlović-Lažetić
Program Chair
DMBI 2012

Conference program

June 26th, Day One

Location: Small Hall, Ilija M. Kolarac Foundation

08:30 - 09:00	Registration and Opening Ceremony
09:00 - 09:30	Representatives of the Ministry, University and Faculty Opening and Welcome Speech
09:30 - 10:00	prof. Saša Malkov, Faculty of Mathematics, University of Belgrade, Serbia *Bioinformatics Research Group at University of Belgrade*

	Morning Session - Chair prof. Gordana Pavlović-Lažetić
10:00 - 10:45	prof. Peter Tompa, VIB Department of Structural Biology, Brussels, Belgium Laboratory of Intrinsically Disordered Proteins, Institute of Enzymology, Budapest, Hungary *Structural Disorder in the Adaptation of Pathogens to Their Hosts*
10:45 - 11:30	prof. Oxana Galzitskaya, Institute of Protein Research RAS, Laboratory of Protein Physics, Pushchino, Russia *Occurrence of Disordered Patterns and Homorepeats in Eukaryotic and Bacterial Proteomes*
11:30 - 11:50	Coffee Break
11:50 - 12:35	dr Marco Punta, Wellcome Trust Sanger Institute, UK *Sequence Conservation in Disordered Regions*
12:35 - 13:20	prof. Zoran Obradović, Centre of Data Analytic and Biomedical Informatics, College of Science and Technology, Temple University, USA *Analysis and Integration of Inconsistent and Unreliable Biomedical Prediction Models*
13:20 - 15:00	Lunch Break

Location: Room 368, School of Physical Chemistry

	Afternoon Session - Chair prof. Peter Tompa
15:00 - 15:45	prof. Zoran Obradović, Centre of Data Analytic and Biomedical Informatics, College of Science and Technology, Temple University, USA *Predictive Modeling of Patient State and Therapy Optimization*
15:45 - 16:30	prof. Nataša Pržulj, Department of Computing, Imperial College London, UK *Network Topology as a Source of Biological Information*
16:30 - 16:45	Coffee Break
16:45 - 17:30	dr Branislava Rakić, Cambridge Cell Networks, UK *Data Management, Text Mining and Data Integration in the SEURAT Cluster*

Mini-course - Location: Room 718, School of Mathematics

17:45 - 19:45	prof. Vladimir Brusić, Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston MA and Health Informatics Group, MET Computer Science, Boston University *Development of Knowledge-Based Systems in Bioinformatics (mini-course)*

June 27th, Day Two

Location: Room 368, School of Physical Chemistry

	Morning Session - Chair prof. Nataša Pržulj
09:00 - 09:45	prof. Vladimir Brusić, Cancer Vaccine Center, Dana-Farber Cancer Institute, Boston MA and Health Informatics Group, MET Computer Science, Boston University *Knowledge-Based Systems for Immune Profiling and Vaccine Discovery*
09:45 - 10:30	prof. Peter Tompa, VIB Department of Structural Biology, Brussels, Belgium Laboratory of Intrinsically Disordered Proteins, Institute of Enzymology, Budapest, Hungary *Structure and Function of Intrinsically Disordered Chaperones*
10:30 - 10:45	Coffee Break
10:45 - 11:30	Jovana Kovačević, Faculty of Mathematics, University of Belgrade, Serbia *Bioinformatics Study of Protein Disorder Content With Respect to its Position and Protein Function*
11:30 - 12:15	Davorka Jandrlić, Faculty of Mechanical Engineering, University of Belgrade, Serbia *T- cell Epitope Frequency in Ordered and Disordered Protein Regions*
12:15 - 12:45	Snack Break
12:45 - 23:00	Excursion

June 28th, Day Three

Location: Room 368, School of Physical Chemistry

	Morning Session - Chair prof. Vladimir Brusić
09:00 - 09:45	prof. Predrag Radivojac,School of Informatics and Computing, Indiana University, USA *Machine Learning Approaches in Protein Function Prediction*
09:45 - 10:30	dr Nevena Veljković,Center for Multidisciplinary Research and Engineering, Vinča Institute of Nuclear Sciences, Serbia *Unraveling Key Determinants of Protein Function by Fourier Transform Based Method for Sequence Analysis*
10:30 - 10:45	Coffee Break
10:45 - 11:30	prof. Branko Dragović, Institute of Physics, University of Belgrade, Serbia *p-Adic Ultrametricity in the Genetic Code and Bioinformatics*
11:30 - 12:15	prof. Miloje Rakočević, Department of Chemistry, Faculty of Science, University of Niš, Serbia *Physicochemical-Mathematical Unity in the Genetic Code - an Elementary Approach*
12:15 - 12:20	Short Coffee Break
12:20 - 13:05	prof. Goran Nenadić, School of Computer Science and Manchester Interdisciplinary BioCenter, University of Manchester, UK *Contextualising and Exploring Molecular Interactions Through Full-Scale Biomedical Text Mining: From Facts to Contradictions*
13:05 - 15:05	Lunch Break

Location: Room 368, School of Physical Chemistry

	Afternoon Session - Chair prof. Nenad Mitić
15:05 - 15:50	prof. Marko Djordjević, Faculty of Biology, University of Belgrade, Serbia *Transcription Start Site Prediction in Bacteria*
15:50 - 16:35	prof. Nenad Švrakić, School of Medicine, Washington University, St. Louis, Missouri, USA, Institute of Physics, University of Belgrade, Serbia *Manipulating Micro Array Data*
16:35 - 16:50	Coffee Break
16:50 - 17:35	dr Andrija Tomović, Novartis Pharma AG, Modeling & Simulation, Basel, Switzerland *Commercial Perspectives on Bioinformatics*
17:35 - 18:20	dr Damjan Krstajić, Research Centre for Cheminformatics, Belgrade, Serbia *Personalized Medicine and Survival Analysis*
18:20 - 18:50	Closing

Table of Contents

Data management/text mining/data integration in the SEURAT cluster

Gordana Apić

Cell Networks, University of Heidelberg, Germany, and CCNet, Cambridge UK
gordana.apic@camcellnet.com

Abstract. SEURAT is a part of an initiative by European Commission's FP7 Health Program and the COLIPA (European Association of Cosmetics Industry) funded initiative with over sixty partners and six research projects representing the building blocks of a common strategy towards the development of solutions for the replacement of current repeated dose systemic toxicity testing in human safety assessment. Part of the data integration and management efforts are planned on a systems biology framework starting with the classification of the pre-existing knowledge from the literature and expanding with integration of the a number of experimental read-outs created in the project.

Keywords: data management, text mining, data integration

Overview

SEURAT is a part of an initiative by European Commission's FP7 Health Programme and the COLIPA (European Association of Cosmetics Industry) funded initiative with over sixty partners and six research projects representing the building blocks of a common strategy towards the development of solutions for the replacement of current repeated dose systemic toxicity testing in human safety assessment. Part of the data integration and management efforts are planned on a systems biology framework starting with the classification of the pre-existing knowledge from the literature and expanding with integration of the a number of experimental read- outs created in the project.

The research projects are creating a wide range of readouts and experimental test, from in vitro assays to different-omics data, and this information needs to be brought into a common context in order to help to shed light on to underlying mechanisms of long term systemic toxicities as well as to help create predictive in vitro and in silico models for predicting toxicity of chemicals.

Since the common goal is to find alternatives to animal testing, information and results of in vivo animal tests and findings in humans described in the body of literature over the last several decades becomes an important source of information and knowledge that can help relate in vitro/in silico findings to in vivo findings. We use a simple approach starting with harvesting the legacy information from the literature and providing partners with useful pre-existing knowledge in a form of a simple database. Such knowledge framework with links

to original references has an advantage that it delivers useful information to the partners very early at the start of the project and serves as a starting point for data integration.

Here we describe an iterative strategy for data integration and management starting with a simple wiki based portal for sharing fairly unstructured information, knowledge from the literature and data samples from the very start of a project and using that in order to define standard operating procedures and data formats for data sharing, integration as well as gathering requirements for the users of this information. The goal is that later in the project this starting initiative evolves into structured databases and complex but easy to use systems biology information management resource.

Transcription start site prediction in bacteria

Marko Djordjević

Faculty of Biology, University of Belgrade, Studentski trg 16,
11000 Belgrade, Serbia
dmarko@bio.bg.ac.rs

Abstract

Transcription start sites (TSS) in bacterial genomes are locations where RNA polymerase binds and initiates transcription. Accurate knowledge of TSS is important not only for bioinformatic applications (e.g. gene and operon predictions), but also as the first and the rate limiting step in understanding transcription regulation. TSS predictions is a classical bioinformatics problem, where available methods show poor accuracy. We here approach this problem from a biophysics perspective, in order to develop a more accurate method for TSS prediction. The main idea is to combine accurate alignments of promoter elements ([1]) with a biophysical model of transcription initiation that we recently developed ([2, 3]). In this talk I will discuss both theoretical modeling of transcription initiation, and our recent advances in understanding promoter specificity. I will also present how the modeling and the analyzed sequence specificity are combined in a biophysics based algorithm for TSS detection.

References

1. Djordjevic M.: Redefining Escherichia coli (70) promoter elements: -15 motif as a complement of the -10 motif, J Bacteriol., 193:6305, 2012
2. Djordjevic M, Bundschuh R.: Formation of the open complex by bacterial RNA polymerase–a quantitative model, Biophys J. 94:4233, 2008
3. Djordjevic M.: Efficient transcription initiation in bacteria: an interplay of protein-DNA interaction parameters, submitted to J. Theor. Biol., 2012

p-Adic Ultrametricity in the Genetic Code and Bioinformatics

Branko Dragović

Institute of Physics, University of Belgrade
Pregrevica 118, 11080 Zemun, Belgrade, Serbia
dragovich@ipb.ac.rs

Abstract. To adequately characterize ultrametric structure of the genetic code, we have successfully applied p-adic distance to the space of 64 codons and 20 amino acids. We suggest to use p-adic distance to investigate similarity (nearness) between strings of DNA, RNA and amino acids. We shall give a review of performed research and some prospects to future investigations.

Keywords: genetic code, bioinformatics, data mining, p-adic distance, ultrametricity.

1. Introduction

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

From mathematical point of view, the GC is a mapping of a set of 64 elements onto a set of 21 elements. There is in principle a huge number of possible mappings, but the genetic code is one definite mapping with a few slight modifications. Hence, for modelling of the GC, the main problem is to find the corresponding structure of the space of 64 and 20 (or 21) elements. It will be demonstrated here that the set of 64 codons, and 20 amino acids, has p-adic structure, where $p = 5$ and 2. Detail exposition of p-adic approach to the genetic code is presented in [2–4] (see also [5] for a similar consideration).

Taking into account p-adic distance between constituents of two strings, we shall also introduce modified Hamming distance and consider its possible application in investigation of similarity in bioinformation systems.

2. p-Adic Structure the Genetic Code

This approach is based on the following idea. Codons, which code the same amino acid, should be close in the information sense. To quantify this closeness

(nearness) we should use some distance. Ordinary distance is unappropriate. From insight to the table of the genetic code (see, e.g. Table 1) one can conclude that distribution of codons is like an ultrametric tree and it suggests use of p-adic distance.

Let us first introduce the following subset of natural numbers:

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

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

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

p-Adic distance between two integers x and y is

$$d_p(x\,,y) = |x - y|_p\,. \tag{2}$$

Since p-adic norm is ultrametric, the p-adic distance (2) is also ultrametric, i.e. it satisfies inequality

$$d_p(x\,,y) \leq\ \max\{d_p(x\,,z)\,,d_p(z\,,y)\} \leq d_p(x\,,z) + d_p(z\,,y)\,, \tag{3}$$

where x, y and z are any three integers.

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

$$d_5(a,\,b) = |a_0 + a_1\,5 + a_2\,5^2 - b_0 - b_1\,5 - b_2\,5^2|_5\,, \tag{4}$$

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

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

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

Ultrametric space $\mathcal{C}_5[64]$ can be viewed as 16 quadruplets with respect to the smallest 5-adic distance, i.e. quadruplets contain 4 elements and 5-adic distance

between any two elements within quadruplet is $\frac{1}{25}$. In other words, within each quadruplet, elements have the first two digits equal and third digits are different.

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

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

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

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

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

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

By the above application of 5-adic and 2-adic distances to $\mathcal{C}_5$ [64] codon space we have obtained internal structure of the codon space in the form of doublets. Just this p-adic structure of codon space with doublets corresponds to the vertebrate mitochondrial genetic code. The other (about 20) known versions of the genetic code in living systems can be viewed as slight modifications of this mitochondrial code, presented at Table 1.

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

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

At Table 2 we assigned numbers $x_0x_1 \equiv x_0 + x_1\,5$ to 16 amino acids which are assumed to be present in dinucleotide coding epoch, and $x_0 = 1, 2, 3, 4$ is attached to four late amino acids which were added during trinucleotide coding. Having these 5-adic numbers for amino acids one can consider distance between them, as well as distances between codons and amino acids: there are 23 codon doublets which are at $\frac{1}{25}$ 5-adic distance with the corresponding 15 amino acids (i.e. 16 a.a. without Lysine), i.e. codons within these doublets and related amino acids are at the same 5-adic distance. The other 5 a.a. are at $\frac{1}{5}$ distance with respect to their codon doublets.

3. p-Adically Modified Hamming Distance

Let $a = a_1\,a_2\,\cdots a_n$ and $b = b_1\,b_2\,\cdots b_n$ be two strings of equal length. Hamming distance between these two strings is $d_H(a,b) = \sum_{i=1}^{n} d(a_i, b_i)$, where $d(a_i, b_i) = 0$ if $a_i = b_i$, and $d(a_i, b_i) = 1$ if $a_i \neq b_i$. We introduce p-adically modified Hamming distance in the following way: $d_{pH}(a,b) = \sum_{i=1}^{n} d_p(a_i, b_i)$, where $d_p(a_i, b_i) = |a_i - b_i|_p$ is p-adic distance between numbers a_i and b_i. When $a_i, b_i \in \mathbb{N}$ then $d_p(a_i, b_i) \leq 1$. If also $a_i - b_i \neq 0$ is divisible by p then $d_p(a_i, b_i) < 1$. In the case of strings as parts of DNA, RNA and proteins, this modified distance is finer and should be more appropriate than Hamming distance itself. For example, elements a_i and b_i can be nucleotides, codons and amino acids with above assigned natural numbers, and primes $p = 2$ and $p = 5$.

Acknowledgements

This work is partially supported by the Ministry of Education and Science, Serbia, contracts 173052 and 174012.

References

1. Watson, J. D. and Baker, T. A. and Bell, S. P. and Gann, A. and Levine, M. and Losick, R.: Molecular Biology of the Gene. CSHL Press, Benjamin Cummings, San Francisco. (2004)
2. Dragovich, B. and Dragovich, A.: A p-Adic Model of DNA Sequence and Genetic Code. p-Adic Numbers, Ultrametric Analysis and Applications, Vol. 1, No. 1, 34–41. (2009). [arXiv:q-bio.GN/0607018v1]
3. Dragovich, B. and Dragovich, A.: p-Adic Modelling of the Genome and the Genetic Code. The Computer Journal, Vol. 53, No. 4, 432–442. (2010). [arXiv:0707.3043v1 [q-bio.OT]]
4. Dragovich, B.: p-Adic Structure of the Genetic Code. NeuroQuantology, Vol. 9, No. 4, 716–727. (2011). [arXiv:1202.2353v1 [q-bio.OT]]
5. Khrennikov A. and Kozyrev, S.: Genetic Code on a Diadic Plane. Physica A: Stat. Mech. Appl., Vol. 381, 265–272. (2007). [arXiv:q-bio/0701007]

Occurrence of disordered patterns and homorepeats in eukaryotic and bacterial proteomes

Oxana V. Galzitskaya and Michail Yu. Lobanov

Institute of protein research, Russian academy of sciences, 142290, Pushchino, Moscow Region, Russia
ogalzit@vega.protres.ru

Abstract

We have constructed the clustered Protein Data Bank and obtained clusters of chains of different identity inside each cluster. We have compiled the largest database of disordered patterns (141) from the clustered PDB where identity between chains inside of a cluster is larger or equal to 75simple rules of selection. The results of these analyses would help to further our understanding of the physicochemical and structural determinants of intrinsically disordered regions that serve as molecular recognition elements. We have analyzed the occurrence of the selected patterns in 97 eukaryotic and in 26 bacterial proteomes. The disordered patterns appear more often in eukaryotic than in bacterial proteomes. The matrix of correlation coefficients between numbers of proteins where a disordered pattern from the library of 141 disordered patterns appears at least once in 9 kingdoms of eukaryota and 5 phyla of bacteria have been calculated. As a rule, the correlation coefficients are higher inside of the considered kingdom than between them. The patterns with the frequent occurrence in proteomes have low complexity (PPPPP, GGGGG, EEEED, HHHH, KKKKK, SSTSS, QQQQQP), and the type of patterns vary across different proteomes. The largest fraction of homorepeats of 6 residues belongs to Amoebozoa proteomes (D. discoideum), 46from D. discoideum (Amoebozoa). Homorepeats of some amino acids occur more frequently than others and the type of homorepeats vary across different proteomes. For example, E6 appears most frequent for all considered proteomes for Chordata, Q6 for Arthropoda, S6 for Nematoda. The averaged occurrence of multiple long runs of 6 amino acids in a decreasing order for 97 eukaryotic proteomes is as follows: Q6, S6, A6, G6, N6, E6, P6, T6, D6, K6, L6, H6, R6, F6, V6, I6, Y6, C6, M6, W6, and for 26 bacterial proteomes it is A6, G6, P6, and the others occur seldom. This suggests that such short similar motifs are responsible for common functions for nonhomologous, unrelated proteins from different organisms. A new method (IsUnstruct) based on the Ising model for prediction of disordered residues from protein sequence alone has been developed. The general idea is new and has the distinct advantage over various machine learning methods. For this method we have used the potentials derived from the clustered Protein Data Bank where there are clusters of chains of different identity inside each cluster. For the first time we have added in our method the library of

disordered patterns (141) constructed from the clustered PDB. The IsUnstruct has been compared with other available methods and found to perform well.

T-cell epitope frequency in ordered and disordered protein regions

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

[1] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
nenad@matf.bg.ac.rs
[2] Institute for General and Physical Chemistry, University of Belgrade, Studentski trg 12, 11000 Belgrade, Serbia
mirjanapa@gmail.com
[3] Faculty of Mechanical Engineering, University of Belgrade, 27 marta 16, 11000 Belgrade, Serbia
djandrlic@mas.bg.ac.rs

Abstract. Frequently asked question in immunology is whether the immunodominance of T-cell epitopes is dependent to their localization in the antigen 3D structure or epitope-intrinsic. Using epitope prediction algorithms we have found that both HLA-I and HLA-II epitope frequencies were higher in ordered protein regions, for all analysed taxonomic categories (archaea, bacteria, eukarya, and viridae) and that epitopes appertaining to ordered protein regions were prevalently hydrophobic. Epitope frequency in disordered protein regions of various lengths was constant while in ordered regions had shown a rising trend with prolonging region length. The comparison between predicted and experimentally evaluated epitopes of several tumor associated antigens of cancer/testis antigen group, revealed that majority of epitopes presented by HLA-I and HLA-II molecules were localized in ordered protein regions.

Keywords: T-cell epitopes, disordered/ordered proten regions

1. Introduction

Intrinsically disordered (unstructured) proteins (IDP) ([1]) lack a stable 3D conformation under *in vitro* physiological conditions. Many of them are involved in cell-regulatory functions and cancer ([1]) and are interesting candidates for cancer-vaccine trials. In the last decade, several authors asked whether disorder or order appertaining of the peptide T-cell epitopes ([2]) played the role in immunogenicity. We have compared epitope frequency, binding affinity, average hydrophobicity and location of promiscuous epitopes within predicted ordered and disordered protein regions. The data on predicted and experimentally found epitope location and immunodominance, reported for several tumor-associated antigens, mainly from cancer/testis antigen group (CTA), were correlated.

2. Methods

A database contained 642 proteins from various sources. Most proteins (477) were downloaded from DisProt database (release 4.9, 2009., [3]). For epitope prediction NetMHCpan-2.0 ([4]) and NetMHCIIpan-1.0 ([5]) methods which cover all known HLA-I A, B, C and E allels and HLA-II DRB allels, were used. If there were multiple allels with identical pseudo-sequences, only the first one (alphabeticaly ordered related to allele names) was chosen. For predicting disordered regions VSL2 predictor, variant VSL2B ([6]) was used. On selected number of CTA we have used additional predictors for disorder regions: PONDR VL-XT ([7]) and PONDR-FIT ([3]). The Kyte-Doolittle hydrophobicity scale ([8]) was applied for prediction of hydrophobicity profile. We have developed an application EPDIS (EPItope in DISorder) which integrates all mentioned methods and offers graphical interface (Fig. 1.).

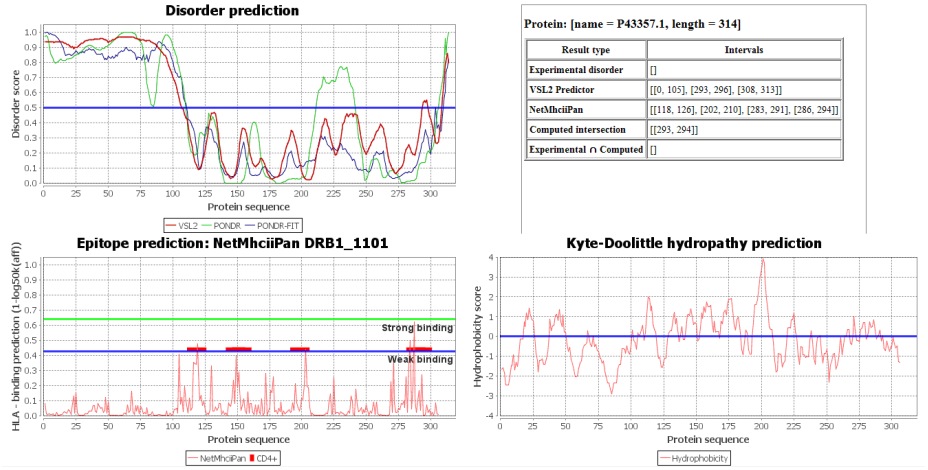

Fig. 1. Human MAGE-A3 protein (UniProt Acc No: P43357.1). Upper left: disorder/order prediction, obtained using VSL-2B, PONDR-VL-XT and PONDR-FIT predictors. Lower left: HLA-II nonamer epitope prediction, using NetMHCpan method for HLA-DRB1_1101 allele. Epitopes, designated with red squares, were experimentally found to induce CD4+ cells in in vitro human studies. Lower right: average hydrophobicity index of nonamer peptides, using Kyte-Doolittle method

3. Results

The number of epitopes (Table 1) in ordered (O) protein regions was 2.84 times higher than in disordered (D) regions for HLA-I allels and 3.60 times higher for HLA-II allels. The same trend (2.64/HLA-I and 3.34/HLA-II) remains after

Table 1. Number of week (WB) and strong (SB)-binding epitopes/100 AA and epitope_in_region hidrophobicity (HER*) in disorder (D), ordered (O) and disorder/order-boundary regions (N) for HLA-I and HLA-II class allels.

HLA	Region type	HLA binding level	Epitopes/100 AA	HER
HLA1	D	SB	88.38	Hydrophlic
		WB	398.19	Hydrophilic
	N	SB	91.05	Hydrophilic
		WB	406.17	Hydrophilic
	O	SB	266.1	Hydrophobic
		WB	1017.83	Hydrophobic
HLA2	D	SB	11.68	Hydrophobic
		WB	201.81	Hydrophilic
	N	SB	15.11	Hydrophobic
		WB	241.02	Hydrophobic
	O	SB	44.04	Hydrophobic
		WB	668.77	Hydrophobic

*If the number of hydrophobic epitopes in region is over 50% the HER is hydrophobic and *vice-versa*

normalization of epitope number on 100AA length. The same conclusion, holds when proteins were grouped according to main taxonomic categories (archaea, bacteria, eukarya, and viridae) (data not shown). The frequency of epitopes (Fig. 2) residing in disordered regions of various lenghts was almost constant (in both bacteria and eukarya), while in ordered regions had shown a rising trend with prolonging region lenght. The epitope frequency growth rate was, however, much slower for O-epitopes in eukarya than in bacteria. From Table 1 it is evident that epitopes, belonging to ordered protein regions are always hydrophobic (for both HLA-I and HLA-II allels). The results are in accordance with HLA-I supertype binding motifs ([9]) and hydrophobic N-terminal positions of the HLA-II class binding peptides ([10]). We have chosen 19 immunogenic cancer-testis antigens, mostly from MAGE-A, NY-ESO and SSX families, that have been intensively studied for cellular immune response and compared epitopes, predicted to be presented by HLA-I and HLA-II antigens and experimentally found ones ([11]), and their localization in ordered and disordered protein regions. Majority of predicted and experimentally-found immunodominant epitopes presented by HLA-I and HLA-II molecules are localized in ordered protein regions, as shown for allele HLA-DRB1_101 for cancer-testis antigen MAGE-A3, Fig. 1 ([12]). In long disordered protein sequences epitopes are frequently flanking the ordered parts of protein or potential disorder-to-order transition elements.

4. Conclusion

Using disorder and epitope prediction algorithms we have found that both HLA-I and HLA-II epitope frequencies were higher in ordered protein regions which may be helpful in mapping potential cancer-vaccine candidate peptides.

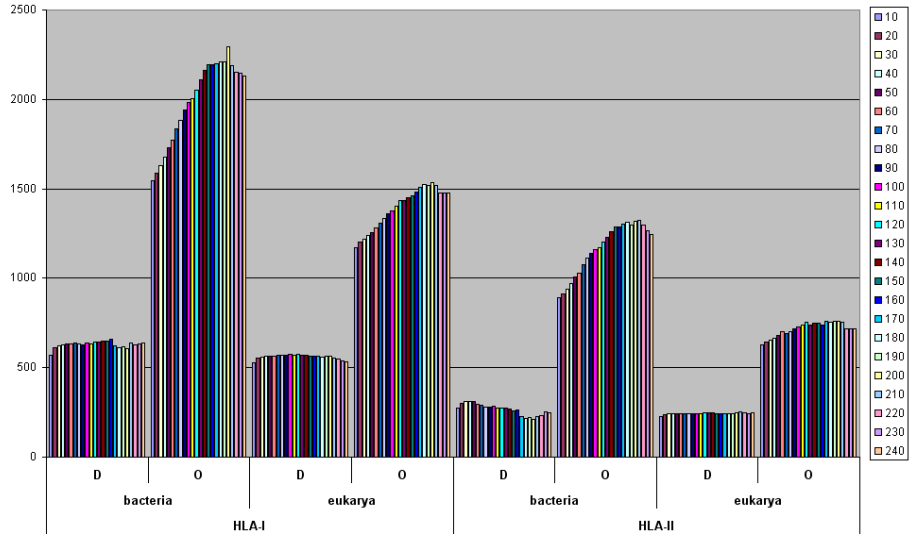

Fig. 2. The number of HLA-I and HLA-II epitopes/100 AA (y-axis), residing in disordered (D) and ordered (O) regions of various length (x-axis) in bacterial and eukaryotic proteins.

Acknowledgements The work presented has been supported by the Ministry of Education and Science, Republic of Serbia, Projects No. 174021. and Project No. 174002.

References

1. Uversky V.N., Dunker A.K.: Understanding protein non-folding, Biochimica et Biophysica Acta - Proteins and Proteomics. 1804 (6) 1231-264 (2010)
2. Carmicle, S. et al.: Antigen three-dimensional structure guides the processing and presentation of helper T-cell epitopes, Molecular Immunology 44 (6) 1159-1168 (2007)
3. Disprot database: http://www.disprot.org
4. NetMhcPan program: http://www.cbs.dtu.dk/services/NetMHCpan
5. NetMhciiPan program: http://www.cbs.dtu.dk/services/NetMHCIIpan-1.0
6. VSL2 predictor: http://www.ist.temple.edu/disprot/predictorVSL2.php
7. Pondr predictor: http://www.pondr.com
8. Kyte, J. and Doolittle, R.: A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology, 157 (1) 105-132 (1982)
9. Sidney, J. et al.: HLA class I supertypes: a revised and updated classification, BMC Immunology, 9:1. (2008)
10. Halling-Brown, et al.: Proteins accessible to immune surveillance show significant T-cell epitope depletion: Implication for vaccine design, Mollecular immunology, 46 (13), 2699-2705. (2009)
11. CTA Database http//www.cancerimmunity.org/peptidedatabase/tumorspecific.htm
12. Consogno, G., et al.: Identification of immunodominant regions among promiscuous HLA-DR-restricted CD4 T-cell epitopes on the tumor antigen MAGE-3, Blood. 101(3), 1038-44. (2003)

Bioinformatics study of protein disorder content with respect to its position and protein function

Jovana Kovačević

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

Abstract. This paper presents a review of research on protein disorder performed by the Bioinformatics Research Group of the Faculty of Mathematics, University of Belgrade. The research has been carried out in two main directions. The first one included a large dataset of around 800 000 prokaryotic proteins where disordered regions were determined by prediction. Disorder content was analyzed with respect to functional categories of proteins, i.e. Clusters of Orthologous Groups of proteins and groups of COGs as well as different genomic, metabolic and ecological characteristics of the organisms that proteins belong to. Second direction involved a small dataset of around 500 prokaryotic, eukaryotic and viral proteins of the DisProt database [http://www.disprot.org] with experimentally determined disordered regions. We investigated the relationship between the disordered content and its position in protein, disordered residues' hydropathy and mass. The sequel of this analysis was searching for correlation and association between disorder fraction of amino acids and their physico-chemical and biochemical characteristics, represented by amino acid indices.

Keywords: disordered proteins / regions, Clusters of Orthologous Groups of proteins, prokaryotes, ecocharacteristics, correlation, association rules, DisProt database, AA indices

1. Introduction

As a result of a growing number of experimental data on protein structure determination, it became evident that a significant number of proteins, under physiological conditions, do not possess a well defined 3D structure. Currently, they are known by different names, with the most frequently used term being "intrinsically disordered proteins (IDP)" and are recently reviewed in detail in [1]. It has been proven that a protein's disorder content is related to its function and that it is often connected to different maladies. This paper presents a survey of work of Bioinformatics Research Group in the field of protein disorder. It includes reviews of two papers, "Bioinformatics analysis of disordered proteins in prokaryotes" [2] and "Computational analysis of position-dependant disorder content in DisProt database" [3], and a summary of one master thesis, "Descriptive models of data mining - applications in bioinformatics" [4].As a result of a growing number of

experimental data on protein structure determination, it became evident that a significant number of proteins, under physiological conditions, do not possess a well defined 3D structure. Currently, they are known by different names, with the most frequently used term being "intrinsically disordered proteins (IDP)" and are recently reviewed in detail in [1]. It has been proven that a protein's disorder content is related to its function and that it is often connected to different maladies. This paper presents a survey of work of Bioinformatics Research Group in the field of protein disorder. It includes reviews of two papers, "Bioinformatics analysis of disordered proteins in prokaryotes" [2] and "Computational analysis of position-dependant disorder content in DisProt database" [3], and a summary of one master thesis, "Descriptive models of data mining - applications in bioinformatics" [4].

2. A research survey of the Bioinformatics Research Group in the field of protein disorder

2.1. Bioinformatics analysis of disordered proteins in prokaryotes

The material for this analysis included proteins from 296 prokaryotic (bacterial and archaeal) completely sequenced genomes. Disordered regions were determined by VSL2B predictor [5]. A thorough analysis of the disorder content of these proteins was carried out with respect to functional categories they belong to, i.e., Clusters of Orthologous Groups of proteins (COGs) and groups of COGs - Cellular processes (Cp), Information storage and processing (Isp), Metabolism (Me) and Poorly characterized (Pc). Disorder content of proteins was also examined from the angle of organisms that proteins belong to, explicitly with respect to various genomic, metabolic and ecological characteristics of the organisms. We used correlations and association rule mining in order to identify the most confident associations between specific modalities of the characteristics considered and disorder content.

The results showed that bacterial proteins from all functional groups except Me have a somewhat higher level of protein disorder than archaeal. Isp and Cp functional groups in particular (specifically COGs L - repair function and N - cell motility and secretion) possess the highest disorder content, while Me proteins, in general, possess the lowest. Fraction of disorder AAs was confirmed to be the lowest for the so-called order-promoting amino acids and the highest for the so-called disorder promoters.

We also investigated which pairs of organism characteristics (e.g., high genome size - high GC content organisms, facultative anaerobic - low GC content organisms, aerobic - high genome size organisms, etc) imply maximum disorder. The results show that archaeal organisms with maximum disorder have the following pairs of characteristics: high GC content - low genome size organisms, high GC content - facultative anaerobic or aquatic or mesophilic organisms, etc. whereas for bacterial organisms these pairs include high GC content - high genome size organisms, high genome size - aerobic organisms, etc.

Minig association rules, we tried to distinguish which organism characteristics are correlated with certain amount of protein disorder (high, medium, low). Most reliable association rules determined relationships between high GC content and high protein disorder, medium GC content and both medium and low protein disorder, anaerobic organisms and medium protein disorder, Gammaproteobacteria and low protein disorder, etc.

Results obtained are well correlated to those previously published, with some extension in the range of disorder level and clear distinction between functional classes of proteins. All the results are published on a web site Prokaryote Disorder Database [http://bioinfo.matf.bg.ac.rs/disorder].

2.2. Computational analysis of position-dependant disorder content in DisProt database

Disordered content from proteins of the DisProt database [6] has been analyzed in respect to the position of AA residues in protein chain. Each protein chain was divided on three parts: N- and C- terminals, both 30 AA long, and the remaining middle part. Fraction of disordered AA residues has been calculated in terminal and middle parts of protein, by protein lengths, by each AA as well as by physico-chemical characteristics.

The motivation for this research was the fact that for a large set of proteins, terminal parts of protein chain are more disordered than its middle part [7]. We tried to check that fact on the proteins of the DisProt database. Fraction of AAs in terminal parts of DisProt proteins is 11.23terminal parts is 17.3N1-10, N11-20, N21-30 and analogously for C-terminal. Fraction of disordered AAs in all tens in terminal parts is similar - around 30whereas in the middle part is around 20fraction of disordered AAs between N21-30 and M part, as well as C21-30 and M part, suggested that the distribution of disordered AA residues in the middle part of protein chains is not uniform. This hypothesis was examined by calculating the fraction of disordered AAs in the middle part divided into subparts 10% long. It was shown that the difference between N21-30 and M, and similarly M and C21-30 is 78middle part is not more than 4%. The result is displayed on Fig. 1.

Calculating the fraction of disorder for all 20 AAs showed that all AAs have higher disorder fraction in terminal parts than in the middle part. We noted a relationship between fractional differences of each AA and its hydropathy. Regarding Kyte-Doolitle scale of hydropathy [8], maximum value of disorder fraction was identified for hydrophobic AAs in all parts.

DisProt proteins contain regions that are marked neither as ordered nor as disordered. We calculated fractional differences between set of disordered regions and set of such, undefined regions and observed that the order of AAs by their fractional difference is almost the same as the order of AAs by their fraction of disorder in the middle part of protein chain (with exception of Y). Comparing this scale with another scale of AA disorder, TOP-IDP scale presented in Campen et al. [9], it is notable that for most AAs the difference between positions of an AA in the two scales is ≤ 3 (except C, H, R, V).

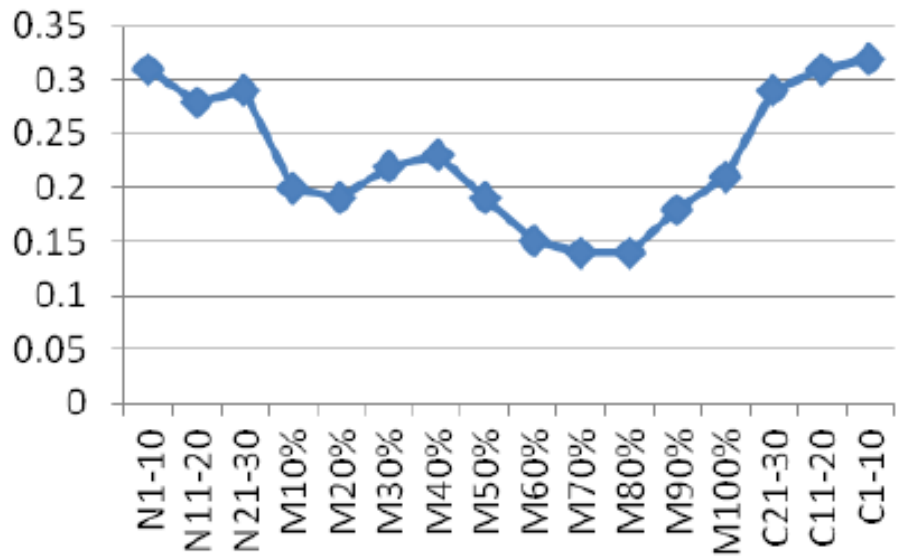

Fig. 1. Fraction of disordered AA residues (y-axis) by protein parts (x-axis).

It was expected that big hydrophobic AAs would be more ordered and small hydrophilic ones more disordered, which implies that big hydrophobic / small hydrophilic AAs would have negative / positive fractional difference between disordered and ordered sets of regions, respectively. In this analysis, we investigated if the same pattern occurs for fractional difference between disordered and undefined sets of regions in DisProt database. Each AA was considered to be small if its mass is less than 114.6 Daltons, which is a median mass for all AAs. Out of 7 hydrophobic AAs, 6 are big and their fractional difference is negative, as expected. From all hydrophilic AAs, only 9 of them is small, and 6 of them have positive fractional difference, as expected.

2.3. Relationship between amino acid disorder fraction and amino acid indices

As an extension to the previous research, in the master thesis "Descriptive models of data mining - applications to bioinformatics" [4], potential causes for the lack of protein structure is considered through association analysis between amino acids disorder fraction (disorder coefficient) and their indices representing physico-chemical and biochemical characteristics. Indices are taken from the amino acid index database [http://www.genome.jp/aaindex, [10], and the disorder fraction is calculated over proteins in the DisProt database. The analysis resulted in a large set of association rules revealing strong connections between disorder coefficients and hydrophobicity and specific energies indices, as well as

amino-acid frequencies in -sheets. These results may contribute to further understanding of disorder regions and improvement of disorder content predictors.

The relationship between disorder fraction and physico-chemical and biochemical characteristics of amino acids has been a subject of further investigation by more sophisticated methods.

Acknowledgements The work presented has been supported by the Ministry of Education and Science, Republic of Serbia, Project No. 174021.

References

1. Tompa P., Fersht A.: Structure and Function of Intrinsically Disordered Proteins. Boca Raton: Chapman and Hall/CRC Taylor and Francis Group (2010)
2. Pavlovic-Lazetic, G., Mitic N., Kovacevic. J., Obradivic Z., Malkov S., Beljanski M.: Bioinformatics analysis of disordered proteins in prokaryotes. BMC Bioinformatics, 12:66 (2011)
3. Kovacevic J.: Computational analysis of position-dependant disorder content in DisProt database. Genomics, Proteomics, Bioinformatics. In press.
4. Stefanovic P. Descriptive models of data mining - applications in bioinformatics, Master thesis, Faculty of Mathematics, University of Belgrade (2012)
5. Peng K., Radivojac P., Vucetic S., Dunker A.K., Obradovic Z.: Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics, 7:208, 1-17. (2006)
6. Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK.: DisProt: the Database of Disordered Proteins. Nucleic Acids Res, 35(Database issue):D786-793. (2007)
7. Lobanov, M., Garbuzynskiy S., Galzitskaya O.: Statistical Analysis of Unstructured Amino Acid Residues in Protein Structures. Biokhimiya, Vol. 75, 236-246 (2009)
8. Kyte, J. and Doolittle, R.: A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157 (1) 105-132 (1982)
9. Campen A., Williams R., Brown C., Meng J., Uversky V., Dunker A.K.: TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. Protein and Pept Lett, 15(9):956-963. (2008)
10. Kawashima S., Pokarowski P., Pokarowska M., Kolinski A., Katayama T., Kanehisa M.: AAIndex: amino acid index database, progress report, Nucleic Acids Research, Volume: 36, Database issue, 202-205 (2008)

Personalized medicine and survival analysis

Damjan Krstajić[1] and Ljubomir Buturović[1]

[1] Research Centre for Cheminformatics, Vidikovacki venac 80v/19,
11030 Belgrade, Serbia
Damjan.Krstajic@rcc.org.rs
[2] Pathwork Diagnostics, 595 Penobscot Drive, Redwood City, CA 94063, USA
lbuturovic@pathworkdx.com

Abstract

Many believe that genome-based medicine, frequently called personalized medicine, is the future of healthcare. It is based on new technologies, like microarrays, which produce data of high complexity. There is increased effort in bioinformatics and statistics to overcome this problem of high dimensionality. The purpose of our talk is two fold. First, we present our work on predicting survival based on Dutch Breast Cancer Dataset (gene expression measurements from 295 women with breast cancer). Second, we would like to point out practical challenges that still exist in this field.

Keywords: : survival analysis, personalized medicine, survivalSVM

Bioinformatics Research Group at the Faculty of Mathematics, University of Belgrade

Saša Malkov[1], Gordana Pavlović-Lažetić[1], Miloš Beljanski[2], Nenad S. Mitić[1], Mirjana D. Pavlović[2], and Jovana Kovačević[1]

[1] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
{smalkov,gordana,nenad,jovana}@matf.bg.ac.rs
[2] Institute for General and Physical Chemistry, University of Belgrade, Studentski trg 12, 11000 Belgrade, Serbia
mbel@matf.bg.ac.rs, mirjanapa@gmail.com

Abstract. Bioinformatics Research Group (BRG) was founded in 2002 from the academic and research staff of the University of Belgrade - Faculty of Mathematics and Institute for General and Physical Chemistry. After ten years of the group existence, the organization of the Data Mining in Bioinformatics - DMBI.2012 is a milestone. The primary goal is to establish wider and stronger connections with researchers in the field and to further improve existing collaborations. Here we present the most significant results of the group.

Keywords: bioinformatics, computer science

1. Introduction

Bioinformatics Research Group (BRG) was founded in 2002 from the academic and research staff of the Faculty of Mathematics - University of Belgrade and Institute for General and Physical Chemistry. The initial resources included good will and readiness to work hard.

In the beginning the main goals were to get introduced to some of the significant areas of bioinformatics research and to recognize the topics which could be researched by the group.

Today the group operates in smaller teams, created according to actual research topics and cover different problems and areas. The group members usually work in multiple teams and on multiple subjects.

The organization of the Data Mining in Bioinformatics - DMBI.2012 is a milestone in the group existence. The primary goal is to establish wider and stronger connections with researchers in the field and to further improve existing collaborations.

2. Bioinformatics Research Group

The BRG works as a part of a larger computer science research group at the Faculty of Mathematics, on fundamental research projects financed by the Min-

istry of Education and Science of Republic of Serbia: No.1858 (2003-2006), No.144030 (2007-2010) and 174021 (2011-2014). Project leader was prof. dr Miodrag Živković, and now is prof. dr Predrag Janičić.

The core of BRG consists of professors and assistants of the Faculty of Mathematics and researchers of the Institute for General and Physical Chemistry: prof. dr Gordana Pavlović-Lažetić, sci.adv. dr Miloš Beljanski, sci assoc. dr Mirjana Pavlović, prof. dr Miodrag Živković, prof. dr Nenad Mitić, prof. dr Sasa Malkov and Jovana Kovacević.

Many researchers from different institutions and graduate students participated in the group's research activities, including: prof. dr Novica Blažić, prof. dr Snežana Zarić, prof. dr Zoran Obradović, dr Ana Simonović, dr Jelena Begović, dr Natasa Golić, dr Branko Jovčić, mr Milena Vujošević-Janicić, mr Jelena Graovac, mr Jelena Hadži-Purić, mr Sana Stojanović, dr Andrija Tomović, Goran Predović, mr Ana Manola, mr Milan Jovanović, mr Ivana Božić, mr Davorka Jandrlić, mr Gorana Dacić, mr Slobodanka Marjanović, mr Milena Šošić, Nevena Petrović, Mirjana Maljković, Aleksandar Zeljić, Milan Todorović, Ulfeta Marovac, Aleksandar Stefanović, Darko Živanović, Mirjana Vuković, Marko Stojicević, Ana Jelović, Milica Knežević, mr Vesna Pajić, Predrag Stefanović, Ana Mijalković, and many others.

3. Results

The main results that the Bioinformatics group has achieved refer to the following tasks:

- Bioinformatics analysis of viral genomes
- Bioinformatics analysis of protein secondary structure
- Bioinformatics analysis of bacterial genomes
- Bioinformatics analysis of protein disorder in prokaryotes
- Reconstruction of genome sequences
- Data mining in bioinformatics
- Text mining in bioinformatics

3.1. Bioinformatics analysis of viral genomes

We performed a comparative analysis of SARS CoV isolates (Severe Acute Respiratory Syndrome), potentially fatal atypical pneumonia. It first appeared in Guangdong province of China in November 2002 and probably originated due to genetic exchange between viruses with different host specificity (mammals, aves). We investigated a dataset of 103 SARS-CoV isolates (101 human patients and 2 Palm Civet isolates).

Our research goal was twofold:

- to establish genome polymorphism (differences in isolate structure) through SNPs, insertions and deletions analysis;
- to Establish variant evolution, through:

- analysis and comparison of genome sequences
- grouping the isolates according to different aspects of sequence similarity, eventually pointing to phylogeny and epidemiological dynamics of SARS CoV.

The results obtained include a new algorithm for sequence similarity, isolate classification according to genome polymorphism and genome haplotypes, mutation analysis through distribution of nucleotides over different distances from SNP sites, type of mutation (transition / transversion, synonymous/non-synonymous) as well as spike (S) protein annotation and analysis [1, 2].

Classification schemes based on genome polymorphism (the level and type of deviation from the "average" isolates) and on genome haplotypes (4 previously known + 5 new loci) are in accordance with possible epidemiological spread, both in space and time, and they fit well into the phylogenetic tree of isolates.

3.2. Bioinformatics analysis of protein secondary structure

Understanding relation of primary and secondary structure of proteins is of great importance. Conformational preferences of amino acids, usually called propensities, are used to predict secondary and tertiary structures of proteins. Individual amino acids show intrinsic propensities towards certain secondary structure types. Moreover, any amino acid in the protein could have the influence on secondary structure type at certain position.

We analyzed the correlations of primary and secondary structures of proteins at the same position in the sequence using data from the Protein Data Bank (PDB) [3].

Our results show clearly how the chemical structure of amino acids plays a major role in determining their preferences for specific secondary structures and enabled us to determine rules for predicting the preference of an amino acid towards a particular secondary structure type based only on the chemical structure of its substituents at the C_β or C_γ atoms [4].

Based on clear preferences of amino acids towards certain secondary structures, we classified amino acids into four groups: $_\alpha$-helix preferrers, strand preferrers, turn and bend preferrers, and others (the group containing *His* and *Cys*, the amino acids showing no clear preference for any secondary structure). Our results show that amino acids in a same group have similar structural characteristics at their C_β and C_γ atoms. All $_\alpha$-helix preferrers have neither polar heteroatoms on C_β and C_γ atoms, nor branching nor aromatic group on the C_β atom. All strand preferrers have aromatic groups or branching on the C_β atom. All turn and bend preferrers have polar heteroatom on C_β or C_γ atoms or do not have a C_β atom at all.

Later we considered correlations of amino acids and particular secondary structures at different relative positions in the sequences. That enabled us to estimate how far the dependence of secondary structure on amino acid is distributed along the chain. We reported on the results of the research in [5].

3.3. Bioinformatics analysis of bacterial genomes

Genomic Islands (GI) are specific regions in many bacterial genomes, that originated by horizontal gene transfer (HGT) between bacteria. GIs may contribute to bacterial adaptability, evolution and pathogenesis. Thus GIs may have different functions such as:

- provide for additional metabolic activities
- enable symbiosis with other organisms
- antibiotic resistance and secretion, etc.

Pathogenicity islands (PAI) are a subset of GIs that contain a variety of virulence factors, providing for specific host recognition, penetration and colonization of the host organism, and the ability to overcome host defense systems. They are characterized by different compositional features, e.g., GC content, as well as different functional features.

Our research goal was to characterize GIs more precisely and to understand them better in order to be able to predict GIs. We applied a linguistic method - exhaustive n-gram analysis of annotated GIs, in addition to other compositional features. The method was trained on the E. coli O157:H7 EDL933 genome, and then used to predict GIs in 14 Enterobacteriaceae family members and in 21 randomly selected bacterial genomes. The results obtained ([6, 7]) included the following:

- Binary sequence classification into GI candidates and others
- Model-based GI prediction
- Model checking against two databases: HGT DB (GIs determined based on compositional features) and PAI DB (PAIs determined based on both compositional and functional features)
- Improvement of prediction precision compared to other compositional methods

3.4. Bioinformatics analysis of protein disorder in prokaryotes

A significant number of proteins have been shown to be intrinsically disordered, meaning that they lack a fixed 3D structure or contain regions that do not posses a well defined 3D structure. It has also been proven that a protein's disorder content is related to its function. We have performed an exhaustive analysis and comparison of the disorder content of proteins from prokaryotic organisms with respect to functional categories they belong to, i.e., Clusters of Orthologous Groups of proteins (COGs) and groups of COGs-Cellular processes (Cp), Information storage and processing (Isp), Metabolism (Me) and Poorly characterized (Pc).

We also analyzed the disorder content of proteins with respect to various genomic, metabolic and ecological characteristics of the organism they belong to. We used correlations and association rule mining in order to identify the

most confident associations between specific modalities of the characteristics considered and disorder content.

Materials used include all the proteins from organisms in the superkingdoms Archaea and Bacteria that contained annotated COGs of proteins (as of November 20th 2009).

Bacteria are shown to have a somewhat higher level of protein disorder than archaea, except for proteins in the Me functional group. It is demonstrated that the Isp and Cp functional groups in particular (L-repair function and N-cell motility and secretion COGs of proteins in specific) possess the highest disorder content, while Me proteins, in general, posses the lowest. Disorder fractions have been confirmed to have the lowest level for the so-called order-promoting amino acids and the highest level for the so-called disorder promoters ([8]).

Some of the most reliable association rules mined establish relationships between high GC content and high protein disorder, anaerobic organisms and medium protein disorder, etc. A web site Prokaryote Disorder Database has been designed and implemented ([9]), which contains complete results of the analysis of protein disorder performed for 296 prokaryotic completely sequenced genomes.

Another part of protein disorder research refers to the DisProt database ([10]). It has been performed with respect to position of disordered residues. Three parts of each protein chain were considered: N- and C- terminals, and the remaining middle part. The results show that, in terminal parts, the overall fraction of amino acid (AA) residues is lower than the fraction of disordered AA residues. Disorder fraction has been analyzed for each of 20 AA in the three parts of proteins with respect to their hydropathy and mass ad a new scale of AAs has been introduced according to their disorder content in the middle part of proteins.

3.5. Reconstruction of genome sequences

Contemporary genome sequencing methods are most often based on shotgun sequencing. Sequences are read in small parts, which are afterwards assembled in longer sequences - contigs. Two significant different tasks are the sequencing of complete genomes and the sequencing of RNA transcriptomes. Genome assembly should result in a small number of long contigs, while transcriptomic assembly provides a larger number of smaller sequences. An assembly may use a reference genome, if there is any. Such assembly is called mapping assembly. Another kind of assembly is De-novo assembly, if assembly cannot use a reference genome.

In collaboration with dr Ana Simonović, from Institute for Biological Research "Sinisa Stanković", we worked on De-novo assembly of Centaurium erythraea Transcriptome. We begun the development of new sequence assembler Sequnomia, using algorithms based on de Bruijn graphs. The development is in a testing phase.

In collaboration with the group (dr Jelena Begović, dr Natasa Golić, dr Branko Jovčić) from Institute of Molecular Genetics and Genetic Engineering we worked on project of assembling genom *L*actobacillus paracasei subsp. paracasei

BGSJ2-8 from contigs. We developed new method for assembling genom from contigs which combine BLAST similarities, GC analysis and Data Mining. With this method we assembled almost 99.74% of genome material.

4. Data mining in bioinformatics

The data mining is present in almost all research activities of the group. Wherever the circumstances allowed, we tried to include the students in application of data mining in the approached problems. The students' activities often had a form of final thesis of their master or magisterium studies, or some practical work on the appropriate subjects on their studies (Bioinformatics, Data Mining). Many different areas were approached. The most significant results in proteomics include:

- Determining the distances between amino-acids in proteins, Nevena Petrović, master thesis;
- Application of data mining techniques in order to determine the correlations of antigene regions with disordered segments of the proteins, mr Davorka Jandrlić, magister thesis;
- Determination of repeat structures in aminoacids (and nucleotide) sequences, Ana Jelović, phd student.

There are even more results in genomic:

- Genoms clustering based on repeats and palindromes, mr Gorana Dacić, magister thesis;
- Claster analysis of bacterial genomic islands, mr Slobodanka Marjanović, magister thesis;
- Application of classification to genom n-gram analysis and determination of genomic islands, mr Milena Šošić, magister thesis;
- Genome assembling based on a set of contigs, Darko Živanović, master thesis;
- Determination of riboswitch sequences positions in genome of Lactobacillus paracasei subsp. paracasei BGSJ2-8, Aleksandar Stefanović, master thesis;
- A new method for determination of elements of gene expression control in bacteria, Marko Stojicević, master thesis.

4.1. Text mining in bioinformatics

The main goal of this research was to extend and complement public databases with data contained in unstructured or semi-structured textual form, such as encyclopedia. A two phase method based on finite state transducers (FST) for information extraction from text was developed and applied ([11–13]). As a textual source we experimented with encyclopedia of microorganisms "Systematic Bacteriology", and as target data we considered genotype / phenotype organism characteristics referred to in the Encyclopedia (genome size, pH and temperature, habitat, oxygen requirement etc).

The results obtained include:

- evaluation of the novel method for extracting encyclopedic information from written sources
- collection of transducers for extracting biological data, applicable to other sources in order to extract new data
- creation of a structured, flexible data resource, which contains phenotype and (some) genotype data for 2412 microbial species
- the resulting database can be used as a reliable complementary resource for genotype-phenotype association research.

5. Conclusion

The first decade of Bioinformatics Research Group was not easy. The beginning steps were rather slow, because we had to learn many things that were completely new to us. But now, after these initial steps, everything looks more promising. Remaining committed to our work, we are confident that we will be much more productive in years to come.

Acknowledgements The work presented has been supported by the Ministry of Education and Science, Republic of Serbia, Project No. 174021.

References

1. Pavlović-Lažetić, G., Mitić, N., Beljanski, M.: Bioinformatics analysis of SARS coronavirus genome polymorphism, BMC Bioinformatics, v.5, 65-78. ISSN: 1471-2105, http://www.thomsonscientific.com/ (2004)
2. Pavlović-Lažetić, G., Mitić, Tomović, A., Pavlović, M., Beljanski, M.: SARS CoV genome polymorphism: bioinformatics study, Genomics, Proteomics, Bioinformatics Journal, Vol. 3, No. 1, 18-35. (2005), Elsevier, ISSN 1672-0229
3. Živković, M., Malkov, S., Zarić, S., Vujosević-Janicić, M., Tomasević, J., Predović, G., Blazić, N., Beljanski, M.V.: Statistical Dependence of Protein Secondary Structure on Amino Acid Bigrams, Chemical Industry & Chemical Engineering Quarterly, 12(1), 82. (2006), http://www.doiserbia.nb.rs/Article.aspx?id=1451-93720601082Z
4. Malkov, S., Živković, S., Beljanski, M., Hall, M., Zarić, S.: A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure, Journal of Molecular Modeling, 14(8):769-775 (2008), DOI: 10.1007/s00894-008-0313-0
5. Malkov, S., Živković, S., Beljanski, Stojanović, S., Zarić, S.: Reexamination of Correlations of Amino Acids with Particular Secondary Structures, The Protein Journal, 28(2):74-86 (2009), DOI: 10.1007/s10930-009-9166-3
6. Mitić, N., Pavlović-Lažetić, G., Beljanski, M.: Could N-gram analysis contribute to genomic island determination? Journal of Biomedical Informatics 41, 936-943. (2008) DOI: 10.1016/j.jbi.2008.03.007, http://dx.doi.org/10.1016/j.jbi.2008.03.007
7. Pavlovic-Lažetić, G., Mitić, N., Beljanski, M.: N-gram characterization of genomic islands in bacterial genomes, Computer Methods and Programs in Biomedicine 93, 241-256. (2009) http://dx.doi.org/10.1016/j.cmpb.2008.10.014
8. Pavlovic-Lažetić, G.M., Mitić, N.S., Kovacević, J.J., Obradović, Z., Malkov, S.N., Beljanski, M.V.: Bioinformatics analysis of disordered proteins in prokaryotes, BMC Bioinformatics (2011), 12:66, DOI:10.1186/1471-2105-12-66

9. Pavlović-Lažetić, G.M., Mitić, N.S., Kovacević, J.J., Obradović, Z., Malkov, S.N., Beljanski, M.V.: Prokaryote Disorder Database, http://bioinfo.matf.bg.ac.rs/disorder (2011)
10. Jovana Kovačević: Computational analysis of position-dependent disorder content in DisProt database, Genomics, Proteomics & Bioinformatics, in press.
11. Pajić, V.: Putting Encyclopaedia Knowledge into Structural Form: Finite State Transducers Approach, Journal of Integrative Bioinformatics, Informations management in der Biotechnologie e.V., Germany, 8(2):164. (2011) ISSN 1613-4516
12. Pajić, V., Pavlovic-Lažetić, G., Pajić, M.: Information Extraction from Semi-structured Resources: A Two-Phase Finite State Transducers Approach, Implementation and Application of Automata, Proceedings of 16th International Conference CIAA, Lecture Notes in Computer Science, Springer Berlin / Heidelberg 282-289. (2011) ISBN 3642222552, 9783642222559
13. Pajić, V., Pavlović-Lažetić, G., Beljanski, M., Brandt, B., Pajić, M.: Towards a Database for Genotype-Phenotype Association Research: Mining Data from Encyclopedia, International Journal of Data Mining and Bioinformatics, Inderscience publishers, (2011), ISSN (Online): 1748-5681, ISSN (Print): 1748-5673, http://www.inderscience.com/browse/index.php?journalID=189&action=coming.

Contextualising and Exploring Molecular Interactions through Full-Scale Biomedical Text Mining: from Facts to Contradictions

Goran Nenadić

[1] School of Computer Science, University of Manchester
Manchester Institute of Biotechnology, Manchester, UK
g.nenadic@manchester.ac.uk
[2] Mathematical Institute, Serbian Academy of Sciences,
Belgrade, Serbia

Abstract

The main archive of life sciences literature (Medline) contains over 18 million references and approximately 2,000 are added to it every day. While the information available in these articles represents a vast source of knowledge, its sheer size also presents challenges to researchers in terms of discovering relevant information. Efforts in biomedical text mining seek to mitigate this problem through systematic extraction of structured data from the full-text literature [8, 10]. Several efforts focus on biomolecular events [**?**,3, 7, 10], which are critical for understanding a diversity of biological processes and functions. We have developed BioContext [4], an integrated text mining system for large-scale extraction and contextualisation of biomolecular events. Events are represented by type (e.g. gene expression, transcription, phosphorylation, binding, regulation) and participants (proteins or events), along with contextual information such as species, anatomical location and whether extracted processes have been reported as speculative or negated (i.e. not taking place, cf. [9]). The results from a number of event ([**?**,7]) and named-entity [2, 5, 6] extraction tools have been merged. Application of our system to 10.9 million Medline abstracts and 234,000 open-access full-text articles from PubMed Central yielded over 36 million mentions representing 11.4 million distinct molecular events, with over 290 000 distinct genes/proteins. Given the large number of publications, it is not surprising that some information is repeated over a number of publications, some with a wide consensus (the same event introduced in several papers). We have also developed a method to find reported interactions that potentially conflict each other. The approach uses rich lexical, syntactic, and semantic features. Overall, 72,314 potentially conflicting molecular interactions have been generated by mining the entire body of accessible biomedical literature. An analysis of 50 potentially conflicting interactions revealed that 32 of them (64to identify biologically interesting but conflicting results. The BioContext pipeline is available for download at http://www.biocontext.org, along with the extracted data, which is available for batch download and online browsing. This resource

should also prove useful to the bioinformatics community for further data mining and processing: in addition to 36 million (contextualised) event mentions, there are 80.0 million gene/protein mentions, 70.9 million species mentions, and 56.6 million anatomical mentions [4].

Keywords: text mining, molecular events, bioinformatics, large-scale processing

References

1. Björne, J., Ginter, F., Pyysalo, S., Tsujii, J. and Salakoski, T.: Complex event extraction at PubMed scale, Bioinformatics, 26, i382-390. (2010).
2. Gerner, M., Nenadic, G. and Bergman, C.M.: LINNAEUS: a species name identification system for biomedical literature, BMC Bioinformatics, 11, 85. (2010).
3. Gerner, M., Nenadic, G. and Bergman, C.M.: An exploration of mining gene expression mentions and their anatomical locations from biomedical text, Proceedings of the BioNLP workshop. Uppsala, Sweden. (2010).
4. Gerner M., Sarafraz F., Bergman C., Nenadic G.: BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. Bioinformatics, doi:10.1093/bioinformatics/bts332 (2012)
5. Hakenberg, J., Gerner, M., Haeussler, M., Solt, I., Plake, C., Schroeder, M., Gonzalez, G., Nenadic, G. and Bergman, C.M.: The GNAT library for local and re-mote gene mention normalization, Bioinformatics, 27, 2769-2771. (2011)
6. Huang, M., Liu, J. and Zhu, X.: GeneTUKit: a software for document-level gene normalization, Bioinformatics, 27, 1032-1033. (2011)
7. Miwa, M., Saetre, R., Kim, J.D. and Tsujii, J.: Event extraction with complex event classification using rich features, J Bioinform Comput Biol, 8, 131-146. (2010)
8. Lu,Z.:PubMed and beyond: a survey of web tools for searching biomedical literature. Database, baq036. (2011)
9. Sarafraz,F. and Nenadic,G.: Using SVMs with the command relation features to identify negated events in biomedical literature, The Workshop on Negation and Speculation in Natural Language Processing. Uppsala, Sweden. (2010)
10. Zhou,D. and He,Y.: Extracting interactions between proteins from the literature, J. Biomed. Inform., 41, 393-407. (2008)

Predictive Modeling of Patient State and Therapy Optimization

Zoran Obradović

Director, Data Analytics and Biomedical Informatics Center, Temple University
zoran@ist.temple.edu

Abstract

This talk will discuss the results of our ongoing project aimed to develop and validate effective predictive modeling technology to achieve the following sepsis treatment related aims on high dimensional and noisy data at a clinically relevant scale:

(1) Personalized sepsis therapy optimization for an individual patient's state improvement;
(2) Early diagnosis of sepsis and accurate detection of change in the state of sepsis, and
(3) Gene expression analysis for sepsis biomarkers identification.

These aims are addressed by developing advanced methods for analysis of temporal dependencies in high dimensional multi-source sepsis related data, which show significant mortality reduction in severe sepsis patients.

This is joint research with my Postdoctoral Associate Dr. Vladan Radosavljevic and my Ph.D. students Mohamed Ghalwash, Dusan Ramljak, Kosta Ristovski and Alexey Uversky.

Biography

Zoran Obradovic, professor of Computer and Information Sciences and the director of the Data Analytics and Biomedical Informatics Center at Temple University in Philadelphia is an internationally recognized leader in data mining and bioinformatics. He has published about 240 articles addressing data mining challenges in health informatics, climate and ecological management, the social sciences, and other domains. Obradovic currently serves as an editorial board member at 8 journals and is the executive editor at the journal on Statistical Analysis and Data Mining which is the official publication of the American Statistical Association (ASA). In years 2013 and 2014 he will chair SIAM International Conference on Data Mining.

Analysis and Integration of Inconsistent and Unreliable Biomedical Prediction Models

Zoran Obradović

Director, Data Analytics and Biomedical Informatics Center, Temple University
zoran@ist.temple.edu

Abstract

In biomedical applications, multiple predictors are often developed for the same problem using multiple training datasets of various qualities. Selecting a single model from such a collection based on accuracy evaluation on a small biased set of annotated data is not necessarily the best strategy when the objective is large scale application of the model. In this talk we will discuss how to address this problem by uncertainty analysis in the reference models and in data. In addition, we will present an iterative algorithm for integrating predictions of multiple models without relying on any annotated data. The proposed solutions will be illustrated on the problem of predicting intrinsic disorder in proteins that lack a stable tertiary structure but still have important biological functions.

This is joint research with my Ph.D. students Mohamed Ghalwash and Ping Zhang.

Biography

Zoran Obradovic, professor of Computer and Information Sciences and the director of the Data Analytics and Biomedical Informatics Center at Temple University in Philadelphia is an internationally recognized leader in data mining and bioinformatics. He has published about 240 articles addressing data mining challenges in health informatics, climate and ecological management, the social sciences, and other domains. Obradovic currently serves as an editorial board member at 8 journals and is the executive editor at the journal on Statistical Analysis and Data Mining which is the official publication of the American Statistical Association (ASA). In years 2013 and 2014 he will chair SIAM International Conference on Data Mining.

Network Topology Complements Sequence as a Source of Biological Information

Nataša Pržulj

Imperial College Londonn, Department of Computing, 180 Queen's Gate, London, SW7 2AZ, UK
natasha@imperial.ac.uk

Abstract

Sequence-based computational approaches have revolutionized biological understanding. However, they can fail to explain some biological phenomena. Since proteins aggregate to perform a function instead of acting in isolation, the connectivity of a protein-protein interaction (PPI) network will provide additional insight into the inner working on the cell, over and above sequences of individual proteins. We argue that sequence and network topology give insights into complementary slices of biological information, which sometimes corroborate each other, but sometimes do not. Hence, the advancement depends on the development of sophisticated graph-theoretic methods for extracting biological knowledge purely from network topology before being integrated with other types of biological data (e.g., sequence). However, dealing with large networks is non-trivial, since many graph-theoretic problems are computationally intractable, so heuristic algorithms are sought.

Analogous to sequence alignments, alignments of biological networks will likely impact biomedical understanding. We introduce a family of topology-based network alignment (NA) algorithms, that we call GRAAL algorithms, which produces by far the most complete alignments of biological networks to date: our alignment of yeast and human PINs demonstrates that even distant species share a surprising amount of PIN topology. We show that both species phylogeny and protein function can be extracted from our topological NA. Furtermore, we demonstrate that the NA quality improves with integration of additional data sources (including sequence) into the alignment algorithm: surprisingly, 77.7% of proteins in the baker's yeast PIN participate in a connected subnetwork that is fully contained in the human PIN suggesting broad similarities in internal cellular wiring across all life on Earth [1]. Also, we demonstrate that topology around cancer and non-cancer genes is different and when integrated with functional genomics data, it successfully predicts new cancer genes in melanogenesis-related pathways; our predictions are phenotypically validated [2]. Finally, we find that aging, cancer, pathogen-interacting, drug-target and genes involved in signaling pathways are topologically "central" in the network, occupying dense network regions and "dominating" other genes in the network [3]. Hence, network topol-

ogy is a valuable source of biological information that can suggest novel drug targets and impact therapeutics.

References

1. O. Kuchaiev and N. Przulj: Integrative Network Alignment Reveals Large Regions of Global Network Similarity in Yeast and Human, Bioinformatics, 27(10): 1390-1396 (2011)
2. T. Milenkovic, V. Memisevic, A. K. Ganesan, and N. Przulj: Systems-level Cancer Gene Identification from Protein Interaction Network Topology Applied to Melanogenesis-related Interaction Networks, Journal of the Royal Society Interface, 7 (44), 423-437, March 6 (2010)
3. T. Milenkovic, V. Memisevic and N. Przulj: Dominating Biological Networks, PLoS ONE, 6(8):e23016 (2011)

Sequence conservation in disordered regions

Jaina Mistry[1], Antonio Deiana[2], Andrea Giansanti[2], Alex Bateman[1], and Marco Punta[1]

[1] Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK
{jm14, agb, mp13}@sanger.ac.uk
[2] Department of Physics, La Sapienza University of Rome, Department of Physics, La Sapienza Ple, A, Moro 5, 00185 Rome, Italy
{Antonio.Deiana, Andrea.Giansanti}@roma1.infn.it

Abstract

Intrinsically disordered regions are predicted to occur in a large number of proteins in all kingdoms of life, with proteins in eukaryotes being the most 'disordered' [1]. Different flavours of disorder have been described, ranging from short flexible loops within well-structured domains, to long unstructured regions.

At Pfam we are interested in detecting disordered regions as a means to better identify boundaries of structured domains in full-length proteins. Indeed, while some intrinsically disordered regions may exhibit conservation, most appear to be highly variable in sequence, making them problematic targets for family building [2, 3].

In this work, we discuss the presence of disorder within the current set of Pfam families. We use the collection of experimentally annotated disordered regions from the DisProt database [4] as well as predictions provided by the method IUPred [5].

Keywords: bioinformatics, protein intrinsic disorder, sequence conservation, Pfam

References

1. Schlessinger, A., Schaefer, C., Vicedo, E., Schmidberger, M., Punta, M. and Rost, B.: Protein disorder–a breakthrough invention of evolution? Curr Opin Struct Biol, 21, 412-418. (2011)
2. Chen, J.W., Romero, P., Uversky, V.N. and Dunker, A.K.: Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J Proteome Res, 5, 879-887 (2006)
3. Bellay, J., Han, S., Michaut, M., Kim, T., Costanzo, M., Andrews, B.J., Boone, C., Bader, G.D., Myers, C.L. and Kim, P.M.: Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biology, 12, R14. (2011)

4. Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N. et al. DisProt: the Database of Disordered Proteins. Nucleic Acids Research, 35, D786-793. (2007)
5. Dosztanyi, Z., Csizmok, V., Tompa, P. and Simon, I.: The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. Journal of Molecular Biology, 347, 827-839. (2005)

Machine learning approaches in protein function prediction

Predrag Radivojac

School of Informatics and Computing
Indiana University, USA
predrag@indiana.edu

Abstract

Understanding protein function is one of the keys to understanding life at the molecular level. It is also important in the context of human disease because many conditions arise as a consequence of alterations of protein function. The recent availability of relatively inexpensive sequencing technology has resulted in thousands of complete or partially sequenced genomes with millions of functionally uncharacterized proteins. Such a large volume of data, combined with the lack of high-throughput experimental assays to functionally annotate proteins, attributes to the growing importance of automated function prediction.

In this talk, I will address state-of-the-art methods for the computational protein function prediction. Such approaches can be broadly classified into (i) those that predict the overall biochemical or biological roles of the molecule, and/or (ii) those that identify specific residues that are necessary for a particular function. Both groups of methods will be generally discussed with emphasis on the role of machine learning as well as on how such methods can be exploited to understand the molecular basis of disease upon mutation.

Biosketch

Predrag Radivojac, Associate Professor of Informatics and Computing, Indiana University - Bloomington. Prof. Radivojac received his Bachelor's and Master's degrees in Electrical Engineering from the University of Novi Sad and University of Belgrade, Serbia. His Ph.D. degree is in Computer Science from Temple University (2003) under the direction of Prof. Zoran Obradovic and co-direction of Prof. Keith Dunker. In 2004 he held a post-doctoral position in Keith Dunker's lab at Indiana University School of Medicine, after which he joined the School of Informatics and Computing, Indiana University - Bloomington. Prof. Radivojac's research interests are in the areas of computational protein function prediction, MS/MS computational proteomics, and machine learning. He received

a National Science Foundation (NSF) CAREER Award in 2007 and his projects are currently supported by NSF and National Institutes of Health (NIH). He is currently an Editorial Board member for the journal Bioinformatics (Oxford University Press) and serves on the Board of Directors of the International Society for Computational Biology (ISCB).

Physicochemical-mathematical unity in the genetic code - an elementary approach

Miloje M. Rakočević

Faculty of Science, University of Niš, Ćirila i Metodija 2 18000 Niš, Serbia
milemirkov@open.telekom.rs
(now retired on the address: Milutina Milankovica 118/25 11070 Belgrade, Serbia)

Abstract. It is shown that in genetic code it exists the unity of a strict stereochemical determinism and the determination by chance. The proofs follow from the relations between four stereochemical and four diversity types of amino acids; all this, through a physicochemical-mathematical unity, valid in the genetic code, just at an elementary level.

Keywords: genetic code, stereochemistry, amino acids, diversity, golden mean, mathematical regularities, cyclic periodic system, atoms

1. Introduction

The key dilemma of a possible understanding of the genetic code is contained in two Crick's standpoints [1]: "the code is universal because it is necessarily the way it is for stereochemical reasons"; or "the allocation of codons to amino acids ... was entirely a matter of 'chance'." In our opinion, one of the reasons why this dilemma is so far not resolved, is the fact that the chemistry almost completely was ignored, especially the elementary chemical facts. This work is concerned just to these facts, considering them important for the analysis of the structure and function of amino acids (AAs) and proteins.

2. Background

For resolving the above dilemma, the works of Vladimir shCherbak about arithmetical regularities that follow the classifications of AAs [2], appear to be significant. Thus, he showed that if the amino acids are splitting into two sets, the four-codon and non-four-codon AAs, then the number of nucleons in 15 of the same "heads" (amino acid functional groups) equals the number of nucleons in 15 very different side chains - in 15 non-four-codon AAs (10 x 111 and 1o x 111). On the other hand, the number of nucleons in eight four-codon AAs is determined by Pythagorean triplet (3-4-5) through multiplication with "Prime quantum 037." For this balance shCherbak says: ‚There is no plausible chemical logic to couple directly the triplets and the amino acids. In other words, the principles of chemistry were not the sought essence of the genetic code" ([2], p. 154). Bearing in mind the said classification of AAs, this conclusion, except

that directly supports the second alternative of Crick, is quite understandable, because it is immediately obvious that the distinction into four-codon and non-four-codon AAs is not chemical, but just formal.

3. Current results

What is however surprising, is the fact that the balances of particles number follow also strict physicochemical distinctions [3], and this result directly supports the first of Crick's alternatives (Fig. 1). By connecting the shCherbak's and our results, we can say that there is a unity of stereochemical determination and the determination by chance.

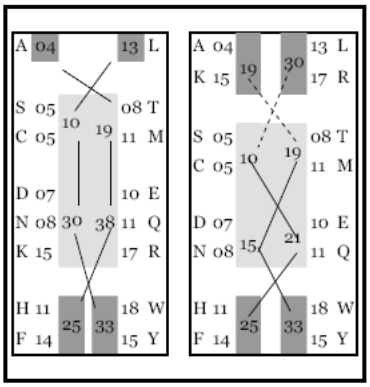

Fig. 1. The atom number within alanine stereochemical type of AAs. The left side: physicochemical hierarchy of amino acid pairs (doublets), determined by left column ("the codogene sequence"); first come aliphatic, then aromatic AAs. At each of both zigzag lines there are 86 atoms. The right side: physicochemical hierarchy of amino acid quartets. On two zigzag lines there are 56 + 30 = 86 and 67 + 19 = 86 atoms, respectively (67 - 56 = 11; notice also that 13 - 4 = 19 -10 = 9 and 33-25 = 38 - 30 = 8)

In support to this, are other our results. Thus, through the determination of golden mean of the first chemically valid order of AAs in standard Genetic Code Table (GCT) - (FL-SP-IMV-TA-YHQ-CWR-NKDE-SRG) [4] - it receives the same four quartets, presented on the right side in Fig. 1 (plus four AAs of three non-alanine stereochemical types: Glycine (G), Prolin (P) and Valine stereochemical type (V-I) [5]; an outcome as a strict Cyclic Invariant Periodic System (CIPS) (Fig. 6 in [6], p. 832). [In central cycle of CIPS there are four chalcogene AAs (ST-CM); follow four contact AAs (GP-VI), then two dicarboxilic AAs and two their amide derivatives (DE-NQ); in next cycle we have two aliphatic AAs and two amine derivatives (AL-KR); finally, come four aromatic AAs (HW-FY).]

At the second chemically valid order of AAs in GCT appear 10 amino acid pairs, five at odd and five at even positions: F-S, L-P, I-T, M-A, V-C, Y-W, H-R, Q-G, N-E, K-D. In such case, odd-even balance exists not only through the number of atoms and nucleons, but also through the number of protons, isotopes, conformations, and through the molecule mass ([7], p. 228).

Furthermore, it will be shown that the four stereochemical types as well as five classes of AAs, presented above, correspond with four types of diversity of AAs, in the arrangement of 2, 4, 6, 8 AAs, respectively (1. G-P, 2. AL-VI, 3. CM-FY-WH, 4. RK-QN-ED-TS), which correspond to the unique arithmetical as well as algebraic regularities [6]. So, the distribution of codons to four types occures through two and two linear equations: x1 y1 = n1,2 and x2 y2 = m1,2, where n1 = 36, m1 = 25, n2= 16, m2 = 9, as the squares of the natural (in a sequence arranged) numbers 6, 5, 4, 3, respectively ([6], p. 829).

It also follows chemicaly valid arrangement of 4 x 5 AAs: [(GP-A-LV, IC-M-FY)/(WH-R-KQ,NE-D-TS)]; then the transformation into 5 x 4 system, through the symmetrical choice, respecting the hierarchy: GSYW, ADMR, Cxxx, Nxxx, and Pxxx, as it is presented in Fig. 3 in [6]. In such arrangement, within first fives AAs there is 26 of atoms; in the second: 26 + 16 = 42; in the third: 42 + 17 = 59 and in the forth fives 59 + 18 = 77. This case is unique within the Table of minimal adding (16+17+18 = 51 as a quarter of 204 = 26 + 42 + 59 + 77) [The Table of minimal adding: first row (01, 02, 03, ... , 10, 11), second row (12, 13, 14, ... , 21, 22) etc. In this Table the number 26 is the lower neighbor of shCherbak's Prime quantum 037 (37 - 26 = 11).]

References

1. Crick, C.H.F.: The Origin of the Genetic Code. J. Mol. Biol. 38, 367-379. (1968)
2. shCherbak, V.: The Arithmetical Origin of the Genetic Code. In: Barbieri, M. (ed.): The Codes of Life, Springer-Verlag, Berlin. (2008)
3. Rakočevic, M.M., Jokic, A.: Four Stereochemical Types of Protein Amino Acids. J. Theoret. Biol. 183, 345-349. (1996)
4. Rakočevic, M.M.: The Genetic Code as a Golden Mean Determined System. Biosystems, Vol. 46, 283-291. (1998)
5. Popov, M.E.: Strukturnaya organizaciya belkov (in Russian). Nauka, Moscow. (1989)
6. Rakočevic, M.M.: Genetic Code as a Coherent System. NeuroQuantology, Vol. 9, No. 4, 821-841. (2011)
7. Rakočevic, M.M.: A Harmonic Structure of the Genetic Code. J. Theoret. Biol. Vol. 229, 221-234. (2004)

Manipulating micro array data

Nenad Švrakić[1,2]

[1] Institute of Physics, University of Belgrade
Pregrevica 118, 11080 Zemun, Belgrade, Serbia
nenads@ipb.ac.rs
[2] Washington University School of Medicine, 660 Euclid, St. Louis, MO, USA

Abstract

Recent development of DNA sequencing and micro array technologies has raised issues of the adequate experimental design in the data collection process, of their organization, classification, and proper handling. Based on the specific needs of biological or medical problem at hand, manipulation of such data has to be carefully planned. We illustrate this by our own work on complex psychiatric disorders where both genetic and endophenotypic data have to be considered in order to extract useful information.

Keywords: bioinformatics, DNA arrays, experimental design, endophenotypes, Word

Commercial Perspectives on Bioinformatics

Andrija Tomović

Novartis International AG, Investor Relations
Novartis Campus, Forum 1,
CH-4056 Basel, Switzerland
andrija.tomovic@novartis.com

Abstract

Bioinformatics has the largest commercial potential within Pharmaceutical and Biotechnological industry. Pharmaceutical research and development is a long and expensive process. Bioinformatics together with all, relatively new, omics technologies/fields have been providing significant contributions to the R&D activities with the ultimate goal to advance drug discovery and development. Nowadays bioinformatics plays an important role in almost all phases of pharmaceutical R&D process. Bioinformatics contributes to the many novel drug discovery paradigms, such as pathway analysis, drug adverse effect predictions, selecting the right dose and many others. Despite all advantages which bioinformatics and other novel technologies have been providing, the number of new molecules brought to market by the biotech and pharmaceutical industry has fallen. Investors and financial markets have started to criticize pharmaceutical industry for too high capital allocation on R&D. Indeed, many different figures imply that R&D productivity has fallen. This trend has been seen as a science problem in general. Can bioinformatics and many other novel innovative disciplines/technologies improve the R&D productivity?

Keywords: bioinformatics, pharmaceutical industry, drug discovery and development

Structural disorder in the adaptation of pathogens to their hosts

Peter Tompa[1,2]

[1] VIB Department of Structural Biology, Vrije Universiteit Brussle, Brussels, Belgium
peter.tompa@vib-vub.be, tompa@enzim.hu
[2] Institute of Enzymology, Hungarain Academy of Sciences, Budapest, Hungary

Abstract

Structural disorder correlates with regulatory and signaling functions, due to which its frequency is high in eukaryotes. It is thought to be also very high in viruses, because i) disorder is compatible with dual coding and high (higher) functional density, ii) viruses probably use disorder to deregulate signaling of the host, and iii) disorder enables rapid evolutionary changes instrumental in evading host defense. Here, results of a few recent studies will be presented and discussed, which advance our understanding in these regards.

In addressing dual coding and structural disorder, we have collected 75 human proteins, which have overlapping coding regions. We found an elevated structural disorder in the dual-coding regions. We also addressed the level of disorder in bacteria of different optimal growth temperature conditions. We found that disorder is an adaptive trait that can change abruptly in evolution: in bacteria living in extreme conditions, it is compromised for sake of adaptation. We have also addressed the level of disorder in all sequenced eukaryotes, thought to be much higher in general than in prokaryotes. We found that structural disorder varies widely in eukaryotes, with the highest levels found in host-changing parasites, whereas the lowest levels are found in obligatory symbionts. These

results relate to structural disorder in viruses because it shows that disorder is an evolutionary trait that has the potential to rapidly change, and to provide an effective way of adaptation to their host. Our studies also raise the interesting question as to the level of disorder in viruses that specialize for eukaryotes of widely different levels of disorder themselves.

Structure and function of intrinsically disordered chaperones

Peter Tompa[1,2]

[1] VIB Department of Structural Biology, Vrije Universiteit Brussle, Brussels, Belgium
peter.tompa@vib-vub.be, tompa@enzim.hu
[2] Institute of Enzymology, Hungarain Academy of Sciences, Budapest, Hungary

Abstract

Structurally disordered proteins (IDPs) can be classified into six functional categories. Based mainly on bioinformatic analysis, we suggested that disordered regions of traditional chaperones or even fully disordered proteins can have potent chaperone activity, probably by an "entropy transfer" mechanism. To carry out detailed structure-function analysis of this phenomenon, we studied two dehydrins of A. thaliana, ERD10 and 14, and show that they are potent chaperones in vitro. Similar observations were made on other IDPs as well. To address the physiological relevance of this effect, we carried out full NMR resonance assignment of the 185 amino acid-long ERD14. Secondary chemical shift and relaxation data show that this IDP is not fully disordered, but have five short regions of somewhat restricted flexibility. In-cell NMR of ERD14 overexpressed in E. coli shows that three of these regions (conserved K-segments) undergo further ordering. ERD14 provides significant protection to cells against stress conditions elicited by various means, in which deletion of K segments have varied effects (unpublished observations). These observations on IDP function are also put in the general context of the evolution of IDP function, by showing that certain IDP functions do not require long evolution refinement, but may suddenly arise in the cell.

Unraveling key determinants of protein function by Fourier transform based method for sequence analysis

Nevena Veljković

Center for the Multidisciplinary Research, Institute of Nuclear Sciences Vinca,
University of Belgrade, Belgrade, Serbia
nevenav@vin.bg.ac.rs

Abstract

Identification of conserved properties of highly variable biological molecules is important for identification of functionally important domains that can be used as therapeutic targets. Epidemics caused by highly pathogenic influenza viruses (HPIV) are a continuing threat to human health and to the world's economy. The principal concern of public health authorities is when and where some of contemporary influenza strains will acquire 1918 pandemic H1N1 virus characteristics i.e. high pathogenicity and infectivity. We have performed sequence analyses based by informational spectrum method and identified conserved properties of hemagglutinin (HA), the envelope protein, of HPAIV H5N1 and pandemic swine influenza virus H1N1 (pH1N1) sequences. Further, we compared informational and structural properties of the HA from animal and human influenza virus subtypes, which are important for the receptor/virus interaction. These characterizations allowed estimation of functional effects of mutations and identification of therapeutic and diagnostic targets for the prevention and treatment of pH1N1 infection and potential new pandemic H5N1 virus. After publication in BMC Struct Biol. 2009, 28;9:62 and in BMC Struct Biol. 2009, 7;9:21. these predictions were confirmed experimentally in vivo and by virus evolution. Our recent developments aimed to track the H5N1 virus evolution towards possible new pandemic virus will be also presented.

Author Index

Ministry of Education and Science of the Republic of Serbia

1921

POSTAL SAVINGS BANK J.S.C.

ПОШТА СРБИЈЕ

Традиционалан и савремен систем

Савремена Пошта својом разгранатом пословном мрежом са више од 1.500 пословница обезбеђује доступност поштанске услуге сваком грађанину на нивоу који је приближан европским стандардима, како у погледу броја становника по јединици поштанске мреже, тако и у погледу рокова и квалитета доставе. Умрежености пословних јединица, аутоматизација пословања и могућност рада у он-лајн режиму, амбициозни даљи планови развоја већ су је учинили препознатљивим и готово незаменљивим инфраструктурним системом у плановима Републике Србије да што већи део државних сервиса приближи грађанима.

Захваљујући томе, Пошта Србије је и у прошлој, макроекономски веома тешкој години, увећала обим услуга, остварила раст прихода и започела нове стратешке пројекте, значајне не само за трансформацију поштанског саобраћаја већ и за модернизацију инфраструктурног потенцијала земље. Пројекти су пре свега усмерени на подизање регионалних поштанско логистичких центара ради аутоматизације прераде поштанских пошиљки, на развој модерних сервиса од значаја за електронску управу, електронско пословање, за свакодневни живот привреде и грађана.

Поред великог инфраструктурног значаја за функционисање друштва, Пошта представља једну од најперспективнијих и најдинамичнијих привредних грана, снажан генератор развоја индустрије, значајан извор националног дохотка и буџетских прихода, а истовремено и најбрже средство у данашње време комуникације међу људима. Њено стратешко опредељење садржано је у трансформацији традиционалне поште у тржишно оријентисану компанију, која ће функционисати, привређивати и зарађивати.

Традиционални и савремени сервиси Поште Србије доступни су корисницима код куће, на послу, на месту где купују или у самој пошти, и то захваљујући моћном комуникационом окружењу: у један систем повезане су све поште, доставни реони, курирске службе, Интернет, позивни центри и поште под франшизом. Посебне погодности Пошта нуди великим пошиљаоцима, јер развија нове сервисе са ширим асортиманом прилагођених поштанских производа који задовољавају специфичне потребе пословања њихових компанија.

Флексибилна, месту и времену прилагођена организациона, кадровска и техничка решења у Пошти подређена су суштинском циљу - успешном и економски валоризованом праћењу промена у глобалном окружењу. Живимо у времену динамичних промена, умеће је што боље их разумети и предвидети, што није увек једноставно. Међутим, материјални ресурси Поште Србије створени досадашњим радом, као и знањем и искуством запослених, гаранција су да ће се и сва будућа искушења успешно превазићи.


RNIDS
Register of National Internet
Domain Names of Serbia

ЦИП – Каталогизација у публикацији
Народна Библиотека Србије, Београд

CONFERENCE Data Mining in Bioinformatics (2012; Belgrade)

Book of Abstracts of the Conference Data Mining in Bioinformatics, June 26th-28th, Belgrade, Serbia : Nenad Mitić (Ed.). – Belgrade : University of Belgrade - Faculty of Mathematics, , 2012 (Belgrade: TMF). – XIV, 53 str. :graf. prikazi : 24cm

Тираж 100. – Abstracts.