Vol. 43 - No. 1 Special Edition 2021

Biologia Serbica

Department of Biology and Ecology,
Faculty of Sciences, University of Novi Sad, Serbia

Book of Abstracts
Belgrade BioInformatics Conference 2021

21-25 June 2021, Vinča, Serbia

http://ojs.pmf.uns.ac.rs/index.php/dbe_serbica/index

Biologia Serbica

JUNE 2021

Vol. 43 - No. 1 Special Edition

Biologia Serbica is formerly Review of Research, Faculty of Sciences, Biology Series (1971-2004), published by Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Serbia

Aim and Scope

Biologia Serbica (BS) is an international scientific, peer-reviewed, open access journal that publishes original research articles, review articles, and short communications in all areas of biology.

Publisher

Department of Biology and Ecology
Faculty of Sciences
University of Novi Sad
Trg Dositeja Obradovića 2
21000 Novi Sad, Serbia
☎: +381 21 485 26 59
℻: +381 21 450 620
www.dbe.uns.ac.rs

Biologia Serbica is published biannually.

Printed by

"Sajnos"
Momčila Tapavice 2,
21000 Novi Sad
+381/21/499-461, +381/21/499-088
sajnosns@gmail.com

Date od publication: June 2018.

Subscription rates

Instructions available at
http://ojs.pmf.uns.ac.rs/index.php/dbe_serbica/index

ISSN 2334-6590

Biologia Serbica

Book of Abstracts
Belgrade BioInformatics Conference 2021

Volume 43 - No. 1
(Special Edition)
2021

Department of Biology and Ecology
Faculty of Sciences
University of Novi Sad

Content

International Advisory Committee

Prof. Alessandro Treves, International School for Advanced Studies, Trieste, Italy
Prof. Francesco Pappalardo, Department of Drug and Health Sciences, University of Catania, Italy
Prof. Guanglan Zhang, Health Informatics Lab, Metropolitan College, Boston University, Boston USA
Prof. Kristian Vlahovicek, Department of Biology, Faculty of Science, University of Zagreb, Croatia
Prof. Lou Chitkushev, Health Informatics Lab, Metropolitan College, Boston University, Boston USA
Prof. Oxana Galzitskaya, Institute of Protein Research, Russian Academy of Sciences, Moscow, Russia
Prof. Paul Sorba, Laboratory of Theoretical Physics and CNRS, Annecy, France
Prof. Peter Tompa, VIB Structural Biology Research Center, Flanders Institute for Biotechnology, Belgium
Prof. Predrag Radivojac, Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA
Prof. Sergey Volkov, Bogolyubov Institute for Theoretical Physics, National Academy of Sciences, Kiev, Ukraine
Prof. Silvio Tosatto, BioComputing UP Lab, University of Padua, Italy
Prof. Vladimir Brusic, Li Dak Sum Chair Professor of Computer Science, University of Nottingham Ningbo, China
Prof. Vladimir Uversky, Department of Molecular Medicine, University of South Florida, USA
Prof. Yuriy Orlov, I.M. Sechenov First Moscow State Medical University, Moscow, Russia
Prof. Zoran Obradovic, Computer and Information Sciences Department, Temple University, Pennsylvania, USA
Prof.Zoran Ognjanovic, Mathematical Institute, Serbian Academy of Sciences and Arts, Serbia

International Program Committee

Dr. Alex Bateman, Protein sequence resources, EMBL-EBI, UK
Prof. Alexandre de Brevern, University Paris Diderot, Sorbonne Paris Cite, France
Dr. Branislava Gemovic, VINCA Institute of Nuclear Sciences, University of Belgrade, Serbia
Prof. Branko Dragovich, Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
Prof. Dragan Matic, Faculty of Sciences, Department of Mathematics and Informatics, University of Banja Luka, Bosnia and Herzegovina
Prof. Goran Nenadic, The University of Manchester, Department of Computer Science, UK
Prof. Gordana Pavlovic-Lazetic, Faculty of Mathematics, University of Belgrade, Serbia
Prof. Hong-Yu OU, State Key Laboratory of Microbial Metabolism, Shanghai Jiao Tong University, Shanghai, China
Prof. Marko Đordevic, Faculty of Biology, University of Belgrade, Serbia
Dr. Milan Senćanski, VINCA Institute of Nuclear Sciences, University of Belgrade, Serbia
Dr. Nataša Kovacevic-Grujicic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Prof. Natasa Przulj, Catalan Institution for Research, Spain; Barcelona Supercomputing Center, Spain; University College London, UK
Prof. Nenad Mitic, Faculty of Mathematics, University of Belgrade, Serbia
Dr. Nevena Veljkovic, VINCA Institute of Nuclear Sciences, University of Belgrade, Serbia
Prof. Sasa Malkov, Faculty of Mathematics, University of Belgrade, Serbia
Prof. Sergei Kozyrev, Department of Mathematical Physics, Steklov Mathematical Institute RAS, Moscow, Russia
Prof. Tamas Korcsmaros, Erhlam Institute, Norwich, UK
Dr. Valentina Djordjevic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr. Vesna Pajic, Seven Bridges Genomics, Beograd, Serbia
Prof. Vladimir Babenko, Institute of Cytology and Genetics, Novosibirsk, Russia

Local Organising Committee

Andela Rodic, Faculty of Biology, University of Belgrade
Dr. Branislava Gemovic, VINCA Institute of Nuclear Sciences, University of Belgrade
Prof. Branko Dragovich, Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
Prof. Gordana Pavlovic-Lazetic, Faculty of Mathematics, University of Belgrade
Dr. Ivana Moric, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade
Dr. Jelena Milicevic, VINCA Institute of Nuclear Sciences, University of Belgrade
Prof. Jovan Pesovic, Faculty of Biology, University of Belgrade
Prof. Jovana Kovacevic, Faculty of Mathematics, University of Belgrade
Prof. Marko Djordjevic, Faculty of Biology, University of Belgrade
Dr. Marko Zivanovic, BioIRC, Univercity of Kragujevac
Dr. Milan Dragicevic, Institute for Biological Research Sinisa Stankovic, University of Belgrade
Dr. Milan Sencanski, VINCA Institute of Nuclear Sciences, University of Belgrade
Nada Djordjevic, Faculty of Mathematics, University of Belgrade
Prof. Nenad Filipovic, Faculty of Engineering, University of Kragujevac
Prof. Nenad Mitic, Faculty of Mathematics, University of Belgrade
Dr. Nevena Veljkovic, VINCA Institute of Nuclear Sciences, University of Belgrade
Dr. Radoslav Davidovic, VINCA Institute of Nuclear Sciences, University of Belgrade
Dr. Sanja Glisic, VINCA Institute of Nuclear Sciences, University of Belgrade
Prof. Sasa Malkov, Faculty of Mathematics, University of Belgrade
Tamara Drljaca, VINCA Institute of Nuclear Sciences, University of Belgrade
Dr. Valentina Djordjevic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade
Prof. Vesna Rakic, Faculty of Agriculture, University of Belgrade
Dr. Vladimir Perovic, VINCA Institute of Nuclear Sciences, University of Belgrade
Prof. Zeljko D. Popovic, Faculty of Sciences, University of Novi Sad

PREFACE

The Belgrade Bioinformatics conference is a biennial event since 2016. This year, we are hosting the third **Belgrade Bioinformatics 2021 conference – BelBI2021**. The aim of the BelBI2021 is to provide a forum for exchange of knowledge and new ideas between scientists in the field, as well as, to include young scientists in this process and provide educational opportunities for them. Previously, we had the pleasure to welcome participants in Belgrade, but the situation with COVID-19 pandemia during 2020. made us postpone our conference and hold it this year as a virtual event. Nevertheless, the research presented in the abstracts available in this book is outstanding and it was an honor to edit it.

Several research institutions, faculties and scientific societies from Serbia have joined their forces to organize this international conference focused on different aspects of bioinformatics. Four Universities participated in the organization – Universities of Belgrade, Novi Sad, Niš and Kragujevac. The Conference is organized by the Vinča Institute of Nuclear Sciences – National Institute of the Republic of Serbia, University of Belgrade, as the the main organizer, and Faculty of Mathematics, University of Belgrade, Faculty of Biology, University of Belgrade, Institute of Molecular Genetics and Genetic Engineering University of Belgrade, Mathematical Institute of SASA and Serbian Society for Bioinformatics and Computational Biology, as co-organizing institutions, in cooperation with several other institutions and societies from Serbia.

This Book of Abstracts covers a wide range of various topics in bioinformatics, including Big data analytics, Machine learning in biological data analysis, Biological networks, Data mining methods and their applications in biology and medicine, Protein structure and function prediction, and much more. Special session of the BelBI2021 is dedicated to the bioinformatics in the field of COVID-19 analysis, which emerged as an immensely important research topic during the previous year, and these abstracts are also included in this book.

The Book of Abstract is printed as a Special Issue of Biologia Serbica, a journal published by the Faculty of Sciences, Department of Biology, University of Novi Sad, one of the Conference co-organizes, and thus, I would like to thank Prof. Željko D. Popović, Managing Editor, for all his effort to bring this enormous work successfully to the finish.

I would like to thank all members of the International Advisory, the International Program and the Local Organizing Committees for their efforts and help to make this event successful. Also, on behalf of the Local Organizing Committee, I would like to express my deepest gratitude to all attendees, and especially to all presenters for their interesting and much appreciated talks. In addition, we owe many thanks to the Ministry of Education, Science and Technological Development of the Republic of Serbia, as well as to the International Centre for Genetic Engineering and Biotechnology (ICGEB) that supported attendance of many students and early stage researchers. Also, the Local Organizing Committee is very grateful to all Conference's sponsors and donors, especially Factory World Wide d.o.o. and Seven Bridges Genomics, with the hope that they will be with us for many years to come.

This book contains 101 abstracts of presentations at the third Belgrade Bioinformatics 2021 conference – BelBI2021. Authors from 21 countries from almost all continents will present their work at the conference. There will be six keynote lectures, forty one invited lectures, twenty seven contributed talks and thirty two poster presentations.

Belgrade, June 2021.
Branislava Gemović
On behalf of BelBI2021

Conference program

DAY 1, 21.6.2021.

14.30 – 15.00	OPENING CEREMONY (TRACK 1)	
KEYNOTE SESSION (TRACK 1)		
15.00 – 15.50	James Collins, USA	Harnessing Synthetic Biology and Deep Learning to Fight Pathogens
TRACK 1 **Bioinformatics and Data Mining of Biological Data (BiDMBD)**		
16.00 – 16.30	Predrag Radivojac, USA	On the developing protocols and guidelines for the use of variant pathogenicity predictors in the clinic
16.30 – 17.00	Ron Shamir, Israel	Utilizing multi-omics and medical records to understand cancer
17.00 – 17.10	BREAK	
17.10 – 17.40	Alex Bateman, UK	Structure Predictions Transform Protein Family Classification
17.40 – 18.10	Richard Bonneau, USA	Integrating generative physical and deep learning approaches to navigate the structure-function-sequence triangle
18.10 – 18.25	Ana Jelović, Serbia	RepeatPlus – Program for finding motifs and repeats in data sequences
TRACK 2 **Biomedical Informatics**		
16.00 – 16.30	Guanglan Zhang, USA	Extraction of Immune Epitope Information
16.30 – 17.00	Lou Chitkushev, USA	TANTIGEN 2.0: a bioinformatics platform that supports T-cell epitope based vaccine design
17.00 – 17.10	BREAK	
17.10 – 17.40	Mona Singh, USA	Deciphering cancer genomes
17.40 – 17.55	Sergey N. Volkov, Ukraine	The possible role of hydrogen peroxide molecules in ion beam therapy of cancer cells
17.55 – 18.10	Ninel Miriam Vainshelbaum, Latvia	Polyploidy-induced atavistic regression to unicellularity and developmental bivalent gene activation in the context of carcinogenesis
18.10 – 18.25	Teodora Đikić, Serbia	Imidazolines: In silico off-target fishing in the class A of G protein-coupled receptors

DAY 2, 22.6.2021.

10.10 – 10.50	POSTER SESSION 1 (A)	
10.50 – 11.50	SPONSOR SESSION – SEVEN BRIDGES GENOMICS An open hour chat with Seven Bridges bioinformatics team	
TRACK 1 **Bioinformatics and Data Mining of Biological Data (BiDMBD)**		
11.50 – 12.20	Fran Supek, Spain	Clustered mutation patterns in cancer genomes
12.20 – 12.35	Ivan Skadrić, Serbia	Entropy changes in worldwide collected HIV-1 subtypes A and B sequences
12.35 – 12.50	Milan Kovačević, Serbia	Benchmarking of Short and Structural Variant Calling Solutions available on SevenBridges Platform within GIAB Benchmarking Framework
12.50 – 13.05	Jaroslaw Synak, Poland	Virxicon: a lexicon of viral sequences

13.05 – 13.15	BREAK	
13.15 – 13.45	Nataša Pržulj, Spain	Network data fusion wwand topological analysis identifies the neighbours of viral targets and differentially expressed genes in Covid-19 as drug target candidates
13.45 – 14.00	Nirupma Singh, India	Database Construction and Network Analysis of Host-Pathogen Protein-Protein Interactions Involved in Microbial CVDs

TRACK 2
Bioinformatics of COVID-19

11.50 – 12.20	Francesco Pappalardo, Italy	Can modeling and simulation help in COVID-19 vaccine research?
12.20 – 12.50	Branka Zukić, Serbia	Precision medicine and COVID-19: importance of host genome profiling and bioinformatics
12.50 – 13.00	BREAK	
13.00 – 13.30	Anđela Rodić, Serbia	Biophysical and bioinformatics approach to study socio-demographic and weather impacts on the SARS-CoV-2 virus transmissibility
13.30 – 13.45	Sofija Marković, Serbia	A bioinformatics approach to inferring environmental drivers of SARS-CoV-2 transmissibility
13.45 – 14.00	Bojana Ilić, Serbia	Understanding Infection Progression under Strong Control Measures through Universal COVID-19 Growth Signatures
LUNCH BREAK		
KEYNOTE SESSION (TRACK 1)		
15.00 – 15.50	Eugene Koonin, USA	The world of viruses, its global organization and evolution

TRACK 1
Bioinformatics and Data Mining of Biological Data (BiDMBD)

16.00 – 16.30	Vladimir Uversky, USA	Unusual Biophysics and Strange Biology of Intrinsic Disorder
16.30 – 17.00	Damiano Piovesan, Italy	Critical assessment of protein intrinsic disorder prediction
17.00 – 17.10	BREAK	
17.10 – 17.40	Frederick (Fritz) Roth, Canada	Improved pathogenicity prediction for rare human missense variants
17.40 – 17.55	Marija Vidović, Serbia	Comparative De Novo Transcriptomic Analysis of Photosynthetically Active and Non-Photosynthetically Active Tissues of Variegated Pelargonium zonale Leaves
17.55 – 18.10	Ana Pantelić, Serbia	De Novo Transcriptome Sequencing of Ramonda serbica: Identification of Late Embryogenesis Abundant Proteins
18.10 – 18.25	Haitham G. Abo-Al-Ela, Egypt	The emerging regulatory roles of noncoding RNAs in immune function of fish: microRNAs versus long noncoding RNAs

TRACK 2
Bioinformatics of COVID-19

16.00 – 16.30	Bjorn Gruning, Germany	No more business as usual: Agile and effective responses to emerging pathogen threats require open data and open analytics.
16.30 – 17.00	Bosiljka Tadić, Slovenia	Assessing the impact of asymptomatic virus carriers by agent-based modelling approach
17.00 – 17.15	Mariem Ghoula, France	Structural analysis of the interaction between the SARS-CoV-2 Spike protein and the human ACE2 receptor

DAY 3, 23.6.2021.

10.10 – 10.50	POSTER SESSION 1 (B)	
KEYNOTE SESSION (TRACK 1)		
10.50 – 11.40	Jörg Menche, Austria	Network Medicine — From protein-protein to human-machine interactions
TRACK 1 **Theoretical Approaches to BioInformation Systems (TABIS)**		
11.50 – 12.20	Sergei Kozyrev, Russia	Genome as a program
12.20 – 12.50	Branko Dragović, Serbia	Biological information and the genetic code
12.50 – 13.00	BREAK	
13.00 – 13.20	Dragana Bajić, Serbia	Clustering as a Support Technique in Phenotyping and Genotyping
13.20 – 13.40	Mirjana M. Maljković, Serbia	Models for Prediction of Structural Alphabet Protein Blocks
LUNCH BREAK		
TRACK 1 **Theoretical Approaches to BioInformation Systems (TABIS)**		
15.00 – 15.30	Konstantin Severinov, USA	Long-Term Persistence of Plasmids Targeted by CRISPR Interference in Bacterial Populations
15.30 – 16.00	Jane Kondev, USA	Action at a Distance in the Yeast Nucleus
16.00 – 16.10	BREAK	
16.10 – 16.40	Ana Conesa, USA	Challenges and solutions in for the integrative analysis of multi-omics data
16.40 – 17.10	Nediljko Budiša, Germany/Canada	Structure and Evolution of the Genetic Code from the Perspective of a Sceptical Experimental Biochemist
TRACK 2 **GALAXY TRAINING**		
12.00 – 18.00	GALAXY TRAINING	
	Introduction to Galaxy	
	Introduction to NGS	
	Introduction to RNA-Seq analysis with Galaxy and R	
	Visualisation	

DAY 4, 24.6.2021.

KEYNOTE SESSION (TRACK 1)		
10.50 – 11.40	Henrik Nielsen, Denmark	Signal peptide prediction: from plain neural networks to language models
TRACK 1 **Bioinformatics and Data Mining of Biological Data (BiDMBD)**		
11.50 – 12.20	Alexandre de Brevern, France	Impact of protein dynamics on secondary structure prediction
12.20 – 12.50	Peter Tompa, Belgium	Biomolecular condensation: macromolecular assembly in the cell by liquid–liquid phase separation and beyond
12.50 – 13.00	BREAK	
13.00 – 13.30	Yuriy Orlov, Russia	Reconstruction of Gene Networks Associated with Complex Disorders on Example of Parkinson Disease
13.30 – 13.45	Vladimir Babenko, Russia	Altered splicing profile of Ptbp1 drives global splicing tune-up upon neuroinflammation response in rat model
13.45 – 14.00	Parakh Sehgal, India	A Subtractive Proteomics Approach to Identify Putative Drug Targets Against Parasitic Species
TRACK 2 **Biomedical Informatics**		
11.50 – 12.20	Goran Nenadić, UK	Linking Genes, Chemicals and Herbs: an Integrated Knowledge Graph from the Stroke Literature and Databases
12.20 – 12.50	Vladimir Babenko, Russia	cAMP mediated genes profiles underscore drastic reduction of dopamine intake along with opioids in the dorsal striatum of fighting deprived aggressive mice
12.50 – 13.00	BREAK	
13.00 – 13.30	Silvano Piazza, Italy	TBA
13.30 – 13.45	Lana Radenković, Serbia	Predicting suicide: serotonin presynapse dynamic modelling and machine learning approach
13.45 – 14.00	Vladimir Perović, Serbia	Innovative bioinformatic approach to kidney transplant wait-list management in the Republic of Serbia
LUNCH BREAK		
KEYNOTE SESSION (TRACK 1)		
15.00 – 15.50	Igor Jurisica	AI is Not Enough: Explainable Biology for Improved Therapies
15.50 – 16.50	POSTER SESSION 2 (A)	
TOURISTIC HOURS (Museum of Yugoslavia, Touristic Organization of Belgrade)		
17.00 – 19.00		

DAY 5, 25.6.2021.

TRACK 1
Bioinformatics and Data Mining of Biological Data (BiDMBD)

10.50 – 11.20	Jiangning Song, Australia	Leveraging the power of data-driven machine learning techniques to address significant biomedical classification problems
11.20 – 11.50	Vladimir Brusić, China	Supervised Machine Learning Techniques for Single Cell RNA-seq: New Opportunities for Clinical Medicine
11.50 – 12.00	BREAK	
12.00 – 12.30	Shanfeng Zhu, China	Recent Advances in Large-scale Protein Function Prediction
12.30 – 12.45	Stefan Spalević, Serbia	Hierarchical Protein Function Prediction with Tail-GNNs
12.45 – 13.00	Jelisaveta Ilić, Serbia	SB MultiCNV: novel method for copy number variations consensus calling

TRACK 2
Biomedical Informatics

10.50 – 11.20	Hong-Yu OU, China	Identification of the conjugative transfer modules of antibiotic-resistant plasmids
11.20 – 11.50	Patrick Aloy, Spain	Extending the small-molecule similarity principle to all levels of biology
11.50 – 12.00	BREAK	
12.00 – 12.30	Mark Wass, UK	Identification of sequence changes in myosin II that adjust muscle contraction velocity
12.30 – 12.45	Staša Stanković, UK	Using human genetics to understand the aetiology of reproductive ageing and its links to later life diseases
12.45 – 13.00	Andrea Mihajlović, Serbia	Inflammatory bowel disease prediction based on metagenomics data
13.00 – 13.15	Teodora Lukić, Serbia	Differential analysis of co-expression networks in the dog mammary gland carcinomas
13.00 – 14.00	POSTER SESSION 2 (B)	

LUNCH BREAK

TRACK 1
Bioinformatics and Data Mining of Biological Data (BiDMBD)

15.00 – 15.30	Nevena Veljković, Serbia	TBA
15.30 – 16.00	Tijana Milenković, USA	Network science reveals a protein's role in aging and a person's risk of mental health problems
16.00 – 16.10	BREAK	
16.10 – 16.40	Gary Bader, Canada	Gene function prediction using unsupervised and semi-supervised biological network integration
16.40 – 16.55	Alexia Sampri, UK	Using simulation studies to compare and evaluate traditional and probabilistic data integration approaches that solve missing variables problem in big biomedical and health datasets

TRACK 2
Biomedical Informatics

15.00 – 15.30	Nataša Milić, Serbia	Meta-analysis of circulating cell-free DNA's role in the prognosis of pancreatic cancer
15.30 – 16.00	Oxana Galzitskaya, Russia	Exploring amyloidogenicity of peptides from ribosomal protein S1 to develop novel AMPs

16.00 – 16.10	BREAK	
16.10 – 16.40	Zoran Obradović, USA	Influence of medical domain knowledge on deep learning for early diagnostics of Alzheimer's disease
16.40 – 16.55	Vladimir M. Jovanović, Germany	E-boxes as ZEB2 binding sites
KEYNOTE SESSION (TRACK 1)		
17.00 – 17.50	Ben Raphael, USA	Quantifying Tumor Heterogeneity using Single-cell and Spatial Sequencing Technologies
17.50 – 18.10	**CLOSING CEREMONY AND POSTER AWARDS (TRACK 1)**	

KEYNOTE SPEAKERS

Quantifying Tumor Heterogeneity using Single-cell and Spatial Sequencing Technologies

Ben Raphael

Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, USA

Abstract:
Tumors are heterogeneous mixtures of normal and cancerous cells with distinct genetic and transcriptional profiles. In this talk, I will present several computational approaches to quantify tumor heterogeneity and reconstruct tumor evolution using data from single-cell DNA and spatial RNA sequencing technologies. For targeted single-cell DNA sequencing, I will describe methods that reconstruct tumor evolution using both somatic single-nucleotide mutations and copy number aberrations, properly accounting for overlap and interactions between these two common classes of mutations. For low-coverage whole-genome single-cell sequencing, I will describe an algorithm CHISEL to compute allele-specific copy numbers and an application of this algorithm to reconstruct tumor evolution for thousands of single cells from a breast tumor. For spatial transcriptomics, I will describe a new algorithm PASTE to align and integrate spatial transcriptomics data from multiple adjacent tissue sections using both transcriptional and spatial similarity. I will illustrate the advantages of multi-section alignment and integration for quantifying heterogeneity in squamous cell carcinomas and inferring cell types in human tissues.

*Corresponding author, e-mail: braphael@cs.princeton.edu

The world of viruses, its global organization and evolution

Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Abstract:
Viruses and virus-like mobile genetic elements are ubiquitous parasites (and sometime symbionts) of all cellular life forms and the most abundant biological entities on earth. The recent, unprecedented advances of comparative genomics and metagenomics have led to the discovery of diverse novel groups of viruses and provide for a vastly improved understanding of the evolutionary relationships within the virosphere. Arguably, we are approaching the point when the global architecture of the virus world can be outlined in its entirety, and the key evolutionary events in each of its domains can be reconstructed.

I will present such an outline of the global organization of the virus world and the corresponding megataxonomy structure that has been recently approved by the International Committee for Taxonomy of Viruses. In particular, I will present the comprehensive evolutionary tree of RNA viruses and discuss the positions of the viruses that cause major human diseases in the different branches of this tree. It is of note that coronaviruses, including SARS-CoV-2, possess the largest and most complex genomes among the RNA viruses known to date, arguably, reflecting the complexity of the virus-host interactions. Although the global structure of the virus world is becoming apparent, major groups of viruses within the established realms, with unique features of genome organization and expression, are being discovered at a high pace. Examples of such new viruses will be presented including the most abundant bacteriophages in the human gut.

*Corresponding author, e-mail: koonin@ncbi.nlm.nih.gov

Signal peptide prediction: from plain neural networks to language models

Henrik Nielsen

Department of Health Technology, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

Abstract:
The signal peptide is the most well-known protein sorting signal, signaling entry into the secretory pathway in all domains of life. Since it was launched in 1996, the program SignalP has been the most popular method for predicting signal peptides from amino acid sequences. In this talk, I will give a historical overview of the various versions of SignalP. Versions 1 to 4 were based on conventional "shallow" feed-forward neural networks, combined with a hidden Markov model in versions 2 and 3. In 2019, SignalP 5.0 was introduced based on deep recurrent neural networks combined with a conditional random field, and it is able to distinguish between several types of signal peptides in Bacteria and Archaea.

In the last part of my talk, I will present a preview of the not yet published SignalP 6.0, which is based on pretrained protein language models. This has made it possible, for the first time, to distinguish between all five known types of signal peptides in prokaryotes, including the rare Tat/SPII and Sec/SPIII types. Furthermore, SignalP 6.0 is able to assign the n-, h-, and c-regions within the signal peptides, enabling detailed analysis of signal peptide properties. In addition, SignalP 6.0 does not need information about the organism of origin, making it well suited for analysis of metagenomic datasets.

*Corresponding author, e-mail: henni@dtu.dk

AI is Not Enough: Explainable Biology for Improved Therapies

Igor Jurisica

University of Toronto, Canada

Abstract:
Integrative computational biology and AI help improving treatment of complex diseases by building explainable models. From systematic data analysis to improved biomarkers, drug mechanism of action, and patient selection, such analyses influence multiple steps of drug discovery pipeline. Data mining, machine learning, graph theory and advanced visualization help characterize interactome and drug orphans with accurate predictions, making disease modeling more comprehensive. Intertwining computational prediction and modeling with biological experiments will lead to more useful findings faster and more economically.

*Corresponding author, e-mail: juris@ai.utoronto.ca

Harnessing Synthetic Biology and Deep Learning to Fight Pathogens

J.J. Collins

Institute for Medical Engineering & Science Department of Biological Engineering
Massachusetts Institute of Technology;
Broad Institute of MIT and Harvard;
Wyss Institute, Harvard University

Abstract:
Synthetic biology is bringing together engineers, physicists and biologists to model, design and construct biological circuits out of proteins, genes and other bits of DNA, and to use these circuits to rewire and reprogram organisms. These re-engineered organisms are going to change our lives in the coming years, leading to cheaper drugs, rapid diagnostic tests, and synthetic probiotics to treat infections and a range of complex diseases. In this talk, we highlight recent efforts that use synthetic biology and deep learning to create novel classes of diagnostics and therapeutics.

*Corresponding author, e-mail: jimjc@mit.edu

Network Medicine — From protein-protein to human-machine interactions

Jörg Menche

Max Perutz Labs, University of Vienna, Vienna, Austria

Abstract:
Integrative computational biology and AI help improving treatment of complex diseases by building explainable models. From systematic data analysis to improved biomarkers, drug mechanism of action, and patient selection, such analyses influence multiple steps of drug discovery pipeline. Data mining, machine learning, graph theory and advanced visualization help characterize interactome and drug orphans with accurate predictions, making disease modeling more comprehensive. Intertwining computational prediction and modeling with biological experiments will lead to more useful findings faster and more economically.

*Corresponding author, e-mail: joerg.menche@univie.ac.at

INVITED SPEAKERS:

Structure Predictions Transform Protein Family Classification

Alex Bateman

Protein sequence resources, EMBL-EBI, UK

Abstract:
Structural prediction models have come of age and are beginning to revolutionise molecular biology. In the recent CASP competition AlphaFold 2 showed accuracies close in many cases to crystal structures. Other methods such as trRosetta and RaptorX still give excellent models that are adequate for many applications. In this talk I will discuss how in collaboration with David Baker's group we released a large collection (>6,300) of structural models for Pfam families. We have begun to dig into this treasure trove to refine, define and classify protein domain families.

Pfam is a collection of over 19,000 protein families with multiple sequence alignments and profile-HMMs. Pfam is widely used to annotate genomes and metagenomes. Ideally Pfam families would correspond to structural domains or repeats found in protein structures. However, often the Pfam family was built before a structure was known. They may be truncated single domains or contain multiple domains. We find that protein structural models can be used to split large multidomain families. Structural models based on Pfam alignments have been less useful to correct truncated domains and these will require the creation of structural models of full length proteins. The structural models have also been instrumental in identifying which superfamilies many Pfam's should belong to. Thus the difference between protein sequence and protein structure classification is becoming smaller and may be unified within the coming few years

*Corresponding author, e-mail: agb@ebi.ac.uk

Impact of protein dynamics on secondary structure prediction

Alexandre G. de Brevern

INSERM UMR_S 1134, DSIMB, Université de Paris, INTS, lab. of excellence GREx,
6, rue Alexandre Cabanel, 75015 Paris, France315100, China

Abstract:
Protein 3D structures support their biological functions. As the number of protein structures is negligible in regards to the number of available protein sequences, prediction methodologies relying only on protein sequences are essential tools. In this field, protein secondary structure prediction (PSSPs) is a mature area, and is considered to have reached a plateau.

Nonetheless, proteins are highly dynamical macromolecules, a property that could impact the PSSP methods. Indeed, in a previous study, the stability of local protein conformations was evaluated demonstrating that some regions easily changed to another type of secondary structure.

The protein sequences of this dataset were used by PSSPs and their results compared to molecular dynamics to investigate their potential impact on the quality of the secondary structure prediction. Interestingly, a direct link is observed between the quality of the prediction and the stability of the assignment to the secondary structure state. The more stable a local protein conformation is, the better the prediction will be. The secondary structure assignment not taken from the crystallized structures but from the conformations observed during the dynamics slightly increase the quality of the secondary structure prediction.

The link between protein dynamics and protein secondary structure prediction was direct. Hence, using the most observed secondary structure assigned state as new assignment (corresponding to a change for 5.4% of the residues) provides an increase of Q3 value from 83.6% to 84.1%. It mainly benefited to residues originally associated to coil regions that are now considered as helical regions (more than half of the cases). This last result underlined the interest to take into account protein flexibility both in the evaluation, but also the assignment.

These results show that evaluation of PSSPs can be done differently, but also that the notion of dynamics can be included in development of PSSPs and other approaches such as de novo approaches.

Keywords:
secondary structure prediction, molecular dynamics, protein structures, B-factors, RMSf, solvent accessibility, DSSP, PSIPRED, structural alphabet, helix, sheet, loop

*Corresponding author, e-mail: alexandre.debrevern@univ-paris-diderot.fr

Biophysical and bioinformatics approach to study socio-demographic and weather impacts on the SARS-CoV-2 virus transmissibility

Igor Salom[1], Andjela Rodic[2*], Ognjen Milicevic[3], Dusan Zigic[1], Bojana Ilic[1], Magdalena Djordjevic[1], Marko Djordjevic[2]

[1] Institute of Physics, University of Belgrade, Pregrevica 118, 11000 Belgrade, Serbia
[2] Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[3] Faculty of Medicine, University of Belgrade, Dr Subotica 8, 11000 Belgrade, Serbia

Abstract

Biophysicists from the Faculty of Biology, in collaboration with scientists and teachers from the Institute of Physics and the Faculty of Medicine, conducted a study of the influence of demographic and climatic factors on the SARS-CoV-2 virus transmission in the population. A nonlinear dynamic compartmental model of epidemic spread was constructed and combined with a bioinformatics approach (collection and analysis of large amounts of data) and the analysis of widespread patterns of infection growth (scaling relations in biophysics). The obtained results indicate that several demographic and meteorological factors significantly affect the basic reproduction number - a measure of the inherent transmission of the virus in a population with given demographic characteristics and weather conditions in the absence of control measures.

The disproportion between the intensive spread of the infection in Wuhan (Hubei) and the much smaller case counts in other Chinese provinces was also analyzed. It has been suggested that this puzzle can be explained by a combination of significantly higher inherent virus transmission in Wuhan and greater effectiveness of epidemic control measures in other provinces.

Overall, the results of these analyzes indicate that the dynamics of epidemic spread may significantly depend on potentially highly heterogeneous and seemingly random factors, such as variations in demographic and meteorological conditions, as well as their complex interaction with introduced control measures. Understanding these factors is crucial, not only for risk estimation during a pandemic but also for long-term prediction of virus behavior in a population if the COVID-19 disease becomes endemic.

Keywords:

compartmental model, bioinformatics, COVID-19, basic reproduction number, environmental effects

*Corresponding author, e-mail: andjela.rodic@bio.bg.ac.rs

Challenges and solutions for the integrative analysis of multi-omics data

Ana Conesa[1,2]

[1] Spanish National Research Council (CSIC), Institute for Integrative Systems Biology (I2SysBio), Spain
[2] Microbiology and Cell Science, University of Florida,USA

Abstract:
Multi-omics approaches have become a reality in both large genomics projects and small laboratories. The analysis of multi-omics data involves multiple steps, from experimental design, preprocessing, statistical analysis, storage, integration, visualization and biological interpretation. Tools are needed to address these steps. Moreover, the multi-omics research community still faces a number of issues that have either not been sufficiently discussed or for which current solutions are still limited. I will present the ConesaLab's bioinformatics tools for multi-omics data analysis and will elaborate on current limitations suggesting points of attention for future research. I will finally discuss new opportunities and challenges brought to the field by the rapid development of single-cell high-throughput molecular technologies.

*Corresponding author, e-mail: ana.conesa@csic.es

Bioinformatics analysis of eukaryotic positively oriented single stranded RNA viruses

Bojana Banović Đeri[1], Dejan Vidanović[2], Bojana Tešović[2], Tamaš Petrović[3], Danijela Ristić[4], Ivan Vučurović[4], Dragana Dudić[5]

[1] *Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia*
[2] *Veterinary Specialized Institute „Kraljevo", Kraljevo, Serbia*
[3] *Scientific Veterinary Institute Novi Sad, Novi Sad, Serbia*
[4] *Institute for Plant Protection and Environment, Belgrade, Serbia*
[5] *Faculty of Informatics, University Union-Nikola Tesla, Belgrade, Serbia*

Abstract
Positively oriented single stranded RNA viruses [ssRNA(+)] persistently affect health and well-being of all eukaryotes, including plants, animals and humans (i.e. SARS-CoV-2, yellow fever, hepatitis C, zika, West Nile, pepper mild mottle virus, etc.). How come these viruses are so wide spread and hard to eradicate? Besides their high changeability, another major reason is their ability to mimic host processes upon entering the host. Only recently it was revealed that ssRNA(+) viruses undergo methylation inside the host in the process that is similar to the methylation of the hosts' own mRNAs. Such process may enable or disable virus to avoid some of the host's defense mechanisms, but it inevitably impacts viral stability and fitness.
Studies on this topic have only started, opening even more questions, with major ones being: how ssRNA(+) methylation, that occurs in the host, impacts viral pathogenicity and are these methylation patterns different in different hosts and for different ssRNA(+) viruses or do these viral methylomes share more universal pattern in concordance with their similar genome organization? Among numerous different methylation patterns of RNA, this research focused on N6-methyladenosine (m6A), as the most common and abundant methylation in eukaryotes, which was confirmed to be present in ssRNA(+) viruses as well.
This study searched for patterns in the primary sequences and secondary structures of ssRNA(+) that are associated to m6A methylation sites relying on the experimentally obtained m6A datasets for eukaryotes and eukaryotic ssRNA(+) viruses. The results are discussed in view of datasets characteristics and study approach.

Keywords:
bioinformatics, m6A, methylome pattern, single stranded RNA viruses, ssRNA(+)

*Corresponding author, e-mail: bbanovicdjeri@gmail.com

Assessing the impact of asymptomatic virus carriers by agent-based modelling approach

Bosiljka Tadic[1*], and Roderick Melnik[2]

[1] *Department of Theoretical Physics, Jozef Stefan Institute, Jamova 39, 1001 Ljubljana, Slovenia*
[2] *Department of Mathematics, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada*

Abstract

The relevance of passive forms of virus transmission in recent SARS-CoV-2 epidemics has been revealed in several studies. In this context, a demanding problem for the control of the epidemic arises from asymptomatic carriers. Through the social mixing occurring at the microscopic interaction scale, they can transmit the viruses to susceptible individuals ending up with potentially life-threatening conditions. The number of asymptomatic hosts and their impact on the infection growth is challenging to assess from the available confirmed cases, resulting in a significant variation from 20% to 80%. Hence, the mathematical modelling of the underlying stochastic processes is of great interest.

Recently, we have developed an agent-based model for latent infection transmissions. In the model, the agent's susceptibility to the virus is its unique feature that differentiates the highly susceptible agents from those that have the susceptibility below a threshold value, and thus, can be asymptomatic. At the same time, it critically determines the agent's role in the infection spreading; the highly susceptible agents can develop severe symptoms and therefore be more infectious, in agreement with some empirical data. The system is driven by the high temporal resolution time series of the social activity and accounts for the virus traceability along the infection path. This modelling framework can effectively differentiate the cases infected by asymptomatic carriers from those infected by symptomatic ones. Consequently, we can assess their relative contributions to the growth of the overall infection curve.

With the extensive simulations comprising eight weeks of evolution time with the hourly resolution, we have shown how the contribution of the asymptomatic carriers varies with the relevant biological parameters that define the critical threshold susceptibility. These are health-related and genetic factors of the population in question and the virus pathogenicity. On the other hand, the overall infection level is primarily determined by the social participation activity and further modified by the changed transmissibility, which the virus mutations can cause.

Keywords

epidemics, SARS-CoV-2, agent-based modeling, asymptomatic

*Corresponding author, e-mail: bosiljka.tadic@ijs.si

Precision medicine and COVID-19: importance of host genome profiling and bioinformatics

Branka Zukić, Nikola Kotur and Biljana Stanković

Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia

Abstract

Clinical picture and course of the disease in patients with COVID-19 vary from asymptomatic to lethal. Precision medicine could discover the cause of this phenomenon by analyzing the individual genomic profiles of the patients.

We aimed to understand a host genetic component of COVID-19 focusing on variants in genes encoding proteases and genes involved in innate immunity, important for susceptibility and resistance to SARS-CoV-2 infection. Also, we wanted to identify phamracogenes and pharmacogenomics markers associated with drugs used for COVID-19 treatment in different clinical protocols in Serbia, and to compare the results with various world populations.

Genotype information of 143 individuals of Serbian origin was extracted from database previously obtained using TruSight One Gene Panel (Illumina). Variants in genes encoding proteases and genes involved in innate immunity were identified and analysed *in silico* (PolyPhen-2, SIFT, MutPred2, Swiss-Pdb Viewer) to predict the impact of the variants to the structure and/or function of proteins. Genotype data from Serbian population was compared with European and 4 super-populations (total 2504 subjects). Data were extracted from VCF files of Phase 3 variant calls of the 1000 Genomes Project (1kGP) sample collection via Ensembl Data Slicer Tool. The level of population genetic variability at each selected loci was examined using the maximal global differences in minor allele frequencies (delta MAF) calculated by subtracting the maximum and the minimum MAF across analyzed population groups, and Fst statistics. Fisher exact test was used to measure differences in genotypes distributions between Serbian and 1kGP populations, applying Bonferoni correction. R software was utilized for genotype data manipulation and statistical calculations.

Based on high alternative allele frequencies in population and the functional effect of the variants, we identified variants in genes encoding proteases and involved in the innate immunity that might be relevant for the host response to SARS-CoV-2 infection. The potential pharmacogenomics markers in pharmacogenes relevant for COVID-19 treatment were also identified. Bioinformatics tools integrated into precision medicine could contribute to better understanding of inter-individual and population-specific genetic susceptibility and resistance to the SARS-CoV-2 infection, therapy response inconsistencies, and could be applied to improve the outcome of the COVID-19 patients.

Keywords:

COVID-19, precision medicine, bioinformatics, host genomics, population pharmacogenomics

*Corresponding author, e-mail: branka.zukic@imgge.bg.ac.rs

Biological information and the genetic code

Branko Dragovich

Institute of Physics, University of Belgrade, Belgrade
Mathematical Institute SASA, Belgrade, Serbia

Abstract
According to the modern scientific developments, the information is getting to be a fundamental notion like space, time and matter. These four fundamental concepts are substantially interconected and represent the basic form of existence of the universe. There is not something without these four characteristics: space, time, matter and information. Being fundamental, there is no complete definition of the information. According to our intuition, we differ what is the information from what it is not. In the present contribution, I consider information as a very special state of the material system with respect to its many possible states. Such very special state of the system determines ifs own evolution and evolution of systems with which it interacts. Depending on complexity of the system one can speak about physical, chemical, biological and other information

Biological information (bioinformation) is related to a special state of a biological system – from viruses to multicellular organisms. The main example of the bioinformation system is DNA, which is a special long sequence of pairs of nucleotides. A part of DNA codes proteins, while the other one is related to the regulation functions. The part that codes codons contains genes whose codons code amino acids, which are building blocks of proteins. The special connection between 64 codons and 20 amino acids with the stop signal is known as the genetic code.

In this talk, I plan to speak about biological information in general and about the genetic code as an illustrative example of bioinformation with its functioning. In particular, I will point out the role of p-adic ultrametrics in description of the genetic code.

Keywords:
bioinformation, the genetic code.

*Corresponding author, e-mail: dragovich@ipb.ac.rs

Critical assessment of protein intrinsic disorder prediction

Damiano Piovesan[1] and Silvio C. E. Tosatto[1]

[1]Dept. Of Biomedical Sciences, University of Padova

Abstract

Intrinsically disordered proteins (IDPs) and regions (IDRs) that do not adopt a fixed, three-dimensional fold under physiological conditions are now well recognized in structural biology. The last two decades have seen an increase in evidence for the involvement of IDPs and IDRs in a variety of essential biological processes and molecular functions that complement those of globular domains.

Identifying unstructured regions of proteins was once part of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) until it was dropped from the challenge following CASP10 in 2012 owing to a lack of new testing and training data, as well as an apparent lack of progress.

In this work, we describe the first edition of CAID, a biennial experiment for the benchmarking of ID and binding predictors on a community-curated dataset of 646 novel proteins obtained from DisProt. CAID has an additional spin as compared to previous benchmarking efforts: it looked at computation time, crucial for large-scale sequence analysis such as metagenomic functional annotation efforts, and it also included a separate contest for binding region prediction within unstructured regions. A total of 43 methods were evaluated, the best methods use deep learning techniques and notably outperform physicochemical methods. The top disorder predictor has F_{max} = 0.483 on the full dataset and F_{max} = 0.792 following filtering out of bona fide structured regions.

The second round of CAID will be synchronized with CASP and will include new sub-challenges such as the prediction of “linker” and “nucleic acid binding” regions.

fIDPnn/lr in the DisProt dataset and AUCpred-np in the DisProt-PDB dataset) leverage evolutionary information, introducing a database search as a preliminary step.

References

Critical assessment of protein intrinsic disorder prediction.
Necci, M., Piovesan, D., CAID Predictors., DisProt Curators., Tosatto, S.C.E.
(2021) Nature Methods.
https://doi.org/10.1038/s41592-021-01117-3

*Corresponding author, e-mail: damiano.piovesan@unipd.it

Clustered mutation patterns in cancer genomes

Fran Supek

Institute for Research in Biomedicine, Parc Científic de Barcelona, C/ Baldiri Reixac 10, 08028 Barcelona, Spain

Abstract

Certain mutagens, including the APOBEC3 (A3) cytosine deaminase enzymes, can create multiple genetic changes in a single event. Activity of A3s results in striking 'mutation showers' occurring near DNA breakpoints; however, less is known about the mechanisms underlying the majority of A3 mutations. We classified the diverse patterns of clustered mutagenesis in tumor genomes, which identified a new A3 pattern: nonrecurrent, diffuse hypermutation (omikli). This mechanism occurs independently of the known focal hypermutation (kataegis), and is associated with activity of the DNA mismatch-repair pathway, which can provide the single-stranded DNA substrate needed by A3, and contributes to a substantial proportion of A3 mutations genome wide. Because mismatch repair is directed towards early-replicating, gene-rich chromosomal domains, A3 mutagenesis has a high propensity to generate impactful mutations, which exceeds that of other common carcinogens such as tobacco smoke and ultraviolet exposure. Cells direct their DNA repair capacity towards more important genomic regions; thus, carcinogens that subvert DNA repair can be remarkably potent.

*Corresponding author, e-mail: fran.supek@irbbarcelona.org

Improved pathogenicity prediction for rare human missense variants

Yingzhou Wu[1], Roujia Li[1], Song Sun[2], Jochen Weile[1], Frederick P. Roth[1]

[1]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada
[2]Sanofi Pasteur, Canada

Abstract

The success of personalized genomic medicine depends on our ability to assess the pathogenicity of rare human variants, including the important class of missense variation. This has been limited by the number and representativity of high-quality examples of rare pathogenic and benign variants for training. Therefore, we developed VARITY, which judiciously exploits a larger reservoir of training examples with uncertain accuracy and representativity, each assigned with a differential weight determined via hyperparameter tuning that optimizes predictive performance on a small high-quality set of annotated variants. VARITY includes new features and feature combinations, while limiting circularity and bias by excluding features informed by variant annotation and protein identity. To provide a rationale for each prediction, we quantify the contribution of features and feature combinations to the pathogenicity inference for each variant. When each is tuned to the same high (90% precision) stringency, VARITY outperforms all previous computational methods evaluated, recovering at least 10% more pathogenic variants.

*Corresponding author, e-mail: fritz.roth@utoronto.ca

Gene function prediction using unsupervised and semi-supervised biological network integration

Gary Bader

The Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Canada

Abstract:
Biological networks constructed from varied data, including protein-protein interactions, gene expression data, and genetic interactions can be used to map cellular function, but each data type has individual limitations such as bias and incompleteness. Unsupervised network integration promises to address these limitations by combining and automatically weighting input information to obtain a more accurate and comprehensive result. However, existing unsupervised network integration methods fail to adequately scale to the number of nodes and networks present in genome-scale data and do not handle partial network overlap. To address these issues, we developed an unsupervised deep learning-based network integration algorithm that incorporates recent advances in reasoning over unstructured data – namely the graph convolutional network (GCN) – and can effectively learn dependencies between any input network, such as those composed of protein-protein interactions, gene co-expression, or genetic interactions. Our method, BIONIC (Biological Network Integration using Convolutions), learns features which contain substantially more functional information compared to existing approaches, linking genes that share diverse functional relationships, including co-complex and shared bioprocess annotation. BIONIC is scalable in both size and quantity of the input networks, making it feasible to integrate numerous networks on the scale of the human genome.

*Corresponding author, e-mail: gary.bader@utoronto.ca

Linking Genes, Chemicals and Herbs: an Integrated Knowledge Graph from the Stroke Literature and Databases

Xi Yang[1,2], Chengkun Wu[1,3], Wei Wang[1], Kai Lu[1], Goran Nenadic[2§]

[1]College of Computer, National University of Defence Technology, Changsha 410073, China
[2]Department of Computer Science, University of Manchester, Manchester M13 9PL, UK
[3]State Key Laboratory of High-Performance Computing, NUDT, Changsha 410073, China

Abstract
Stroke has an acute onset and a high mortality rate, making it one of the most fatal diseases worldwide. Its underlying biology and treatments have been widely studied both in the "Western" biomedicine and the Traditional Chinese Medicine (TCM). However, these two approaches are often studied and reported in insolation, both in the literature and associated databases.

To aid research in finding effective prevention methods and treatments, we integrated knowledge from the literature and a number of databases (e.g. CID, TCMID, ETCM). We employed a suite of biomedical text mining (i.e. named-entity) approaches to identify mentions of genes, diseases, drugs, chemicals, symptoms, Chinese herbs and patent medicines, etc. in a large set of stroke papers from both biomedical and TCM domains. Then, using a combination of a rule-based approach with a pre-trained BioBERT model, we extracted and classified links and relationships among stroke-related entities as expressed in the literature. A knowledge graph, StrokeKG, was then built to integrate information mined from the literature and extracted from the databases.

StrokeKG includes almost 60k nodes of nine types, and 364k links of 30 types, connecting diseases, genes, symptoms, drugs, pathways, herbs, chemical, ingredients and patent medicine. Through manual annotation and linking to the databases, we verified 32k entities and 4,800 relationships in a sub-graph that can be used to facilitate search and improved understanding of this complex disease, for example, by exploring precursor symptoms or pathways for treating related diseases. It can be also used to explore new directions for stroke research and ideas for drug repurposing and discovery.

*Corresponding author, e-mail: G.Nenadic@manchester.ac.uk

Extraction of Immune Epitope Information

Guanglan Zhang

Health Informatics Lab, Metropolitan College, Boston University, Boston, USA

Abstract

Two arms of adaptive immunity, the humoral immunity mediated by B cells and the cell-mediated immunity mediated by T cells, work closely to combat infection and malignancy in an antigen-specific manner. Major histocompatibility complex (MHC) proteins play a vital role in the regulation of cell-mediated immune responses. MHC class I molecules and MCH class II molecules bind short peptides derived from protein antigens through proteolytic mechanisms and subsequently present MHC-peptide complexes on the cell surface for recognition by T cells. Peptides that are presented by MHC and recognized by T-cells are termed T cell epitopes. They are important targets of adaptive immune responses and rational vaccine design. In this talk, we will cover in silico and in vitro methods for immune epitope discovery. In silico methods include databases of immune epitopes, prediction systems for proteasome cleavage, TAP binding, and MHC binding. These in silico tools support downstream in vitro and in vivo studies. Computational tools and databases also play a role in support of developing personalized neoantigen-based cancer vaccines. Neo-antigens arise from tumor-specific mutations, these epitopes are exclusively tumor-specific, and they circumvent T cell tolerance against self-epitopes. Next-generation sequencing and utilization of T cell epitope prediction algorithms provide tools to design cancer vaccines against an individual's own tumor from the Tumor gene sequencing data, one patient at a time. These personalized cancer vaccines offer a leap in cancer treatment in the direction of personalized immunotherapy.

*Corresponding author, e-mail: guanglan@bu.edu

Identification of the conjugative transfer modules of antibiotic-resistant plasmids

Hong-Yu Ou

State Key Laboratory of Microbial Metabolism and School of Life Sciences & Biotechnology, Shanghai Jiao Tong University, 1954 Huashan Road, Shanghai 200030, China

Abstract

Bacterial conjugative plasmids have been highlighted as important vehicles for the dissemination of antibiotic-resistance determinants and pathogenesis. The conjugative transfer regions of the self-transmissible plasmids typically consist of four modules: an origin of transfer (*oriT*) region, relaxase gene, type IV coupling protein (T4CP) gene, and gene cluster for the bacterial type IV secretion system (T4SS) apparatus. In addition, large numbers of mobilizable plasmids, typically carry a limited number of *mob* genes for their DNA processing in conjugation, which is transferable but not self-transmissible. Interestingly, non-conjugative plasmids carrying functional *oriT* sequences can be mobilized by conjugative plasmids. Thus, the identification of these modules in plasmid sequences is important to investigate the self-transfer or mobilizing transfer capability of plasmids.

In this study, we report a web tool, named 'oriTfinder' (https://bioinfo-mml.sjtu.edu.cn/oriTfinder/), as a public resource for in silico detection of conjugative transfer modules in bacterial plasmid sequences, especially in antibiotic resistance plasmids. We first developed a back-end database oriTDB using our collections of known oriT loci, relaxases, and T4CPs of bacterial mobile genetic elements. The oriTfinder then performs rapid homology searches of a query genome sequence against oriTDB. It outputs a simple list and generates a graphic overview of not only the predicted transfer-related functional modules (*oriT* region, relaxase genes, T4CP genes, T4SS gene clusters) but also the extended putative virulence or acquired antibiotic resistance genes. The oriTfinder might facilitate the rapid detection of various conjugative regions in the dynamic plasmids of bacterial pathogens. For example, with the aid of oriTfinder, we elucidated that the emergence of the third-generation cephalosporin-resistant hypervirulent *Klebsiella pneumoniae* resulted from the acquisition of a self-transferable $bla_{\text{DHA-1}}$-carrying plasmid by an ST23 strain.

Keywords:

bacterial plasmid, conjugative transfer, antibiotic-resistant genes, prediction tool, *Klebsiella pneumoniae*

*Corresponding author, e-mail: hyou@sjtu.edu.cn

Action at a Distance in the Yeast Nucleus

Jane Kondev

Martin A. Fisher School of Physics, Brandeis University, USA

Abstract
Chromosomes store genes, but, unlike the hard drives in our computers, their physical properties affect their function as devices for genetic information storage and processing. Cell experiments have revealed that chromosomes in the nucleus assume multiple configurations configurations, and probabilistic models borrowed from polymer physics have been used to describe their shapes. These developments have put front and center the question how fluctuating chromosome shape affects information processing in the cell. In this talk I will describe our experiments on chromosomes in yeast cells, and related theory, which reveal the role of chromosome shape in directing DNA recombination. I will also discuss our recent results on spreading of histone modifications along a chromosome in response to a break in it's DNA. I hope to demonstrate that quantitive experiments and theory can be successfully combined to reveal the physical principles of chromosome-mediated long-range interactions between genes in the cell's nucleus.

*Corresponding author, e-mail: kondev@brandeis.edu

Leveraging the power of data-driven machine learning techniques to address significant biomedical classification problems

Jiangning Song

Monash Biomedicine Discovery Institute, Monash University, Melbourne, Australia

Abstract

Recent advances in high-throughput sequencing have significantly contributed to an ever-increasing gap between the number of gene products ('proteins') whose function is well characterized and those for which there is no functional annotation at all. Experimental techniques to determine the protein function are often expensive and time-consuming. Improving our ability to predict the functional phenotype from genotype is fundamental for understanding the underlying mechanisms of many genetic diseases. Machine-learning (ML) techniques based on statistical learning have recently emerged as efficient solutions to challenging problems of sequence classification, functional annotation or other biomedical classification tasks that were previously regarded difficult to address. In this talk, by combining our recent research works, I will present some important developments in computational algorithms and resources to functionally interpret massive heterogeneous biomedical datasets. In particular, I will highlight three representative research projects to illustrate how biomedical discovery can be alternatively catalysed by data-driven techniques. I will also discuss how ML methods can extract the predictive power from a variety of features that are derived from different aspects of the data and useful strategies that help to contribute to the performance of ML approaches.

*Corresponding author, e-mail: Jiangning.Song@monash.edu

Long-Term Persistence of Plasmids Targeted by CRISPR Interference in Bacterial Populations

Viktor Mamontov[1], Alexander Martynov[1], Natalia Morozova[1,2], Anton Bukatin[3], Dmitry B. Staroverov[4], Konstantin A. Lukyanov[1], Yaroslav Ispolatov[5], Ekaterina Semenova[6], and Konstantin Severinov[1,6,7]

[1] Skolkovo Institute of Science and Technology, Moscow 143028, Russia
[2] Peter the Great St Petersburg State Polytechnic University, St Petersburg 195251, Russia.
[3] Alferov Saint Petersburg National Research Academic University of the Russian Academy of Sciences, 8/3, Khlopina St., 194021 St. Petersburg, Russia
[4] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, Russia
[5] University of Santiago of Chile (USACH), Physics Department, Chile
[6] Waksman Institute for Microbiology, Rutgers, the State University of New Jersey
[7] Institute of Molecular Genetics of National Research Centre "Kurchatov Institute", Moscow 123182, Russia

Abstract
CRISPR-Cas systems provide prokaryotes with an RNA-guided defense against foreign mobile genetic elements (MGEs) such as plasmids and viruses. A common mechanism by which MGEs avoid interference by CRISPR consists of acquisition of escape mutations in regions targeted by CRISPR. Here, using microbiological, live microscopy, and microfluidics analyses we demonstrated that plasmids can persist in *Escherichia coli* cells at conditions of continuous targeting by the type I-E CRISPR-Cas system without acquiring any genetic alterations. We used mathematical modeling to show how plasmid persistence in a subpopulation of cells mounting CRISPR interference is achieved due to the stochastic nature of CRISPR interference and plasmid replication events. We hypothesize that the observed complex dynamics provides bacterial populations with long-term benefits due to the presence of mobile genetic elements in some cells, leading to diversification of phenotypes in the entire community and allowing rapid changes in the population structure to meet the demands of a changing environment.

*Corresponding author, e-mail: severik@waksman.rutgers.edu

TANTIGEN 2.0: a bioinformatics platform that supports T-cell epitope based vaccine design

Lou Chitkushev

Metropolitan College, Boston University, Boston, USA

Abstract

We previously developed TANTIGEN, a comprehensive online database cataloging more than 1,000 T cell epitopes and HLA ligands from 292 tumor antigens. Recently we developed TANTIGEN 2.0, where we significantly expanded coverage in both immune response targets (T cell epitopes and HLA ligands) and tumor antigens. TANTIGEN 2.0catalogs 4,296 antigen variants from 403 unique tumor antigens and more than 1,500 T cell epitopes and HLA ligands. We also included neoantigens, a class of tumor antigens generated through mutations resulting in new amino acid sequences in tumor antigens. TANTIGEN 2.0 contains validated TCR sequences specific for cognate T cell epitopes and tumor antigen gene/mRNA/protein expression information in major human cancers extracted by Human Pathology Atlas. TANTIGEN 2.0 is a rich data resource for tumor antigens and their associated epitopes and neoepitopes. It hosts a set of tailored data analytics tools tightly integrated with the data to form meaningful analysis workflows. It is freely available at http://projects.met-hilab.org/tadb.

*Corresponding author, e-mail: ltc@bu.edu

Identification of sequence changes in myosin II that adjust muscle contraction velocity

Mark Wass

School of Biosciences, University of Kent, Canterbury, UK

Abstract

The speed of muscle contraction is related to body size; muscles in larger species contract at slower rates. Species heart rate is an example of this; a mouse has a heart rate close to 300 beats per minute, while it is only 30 for an elephant. Since contraction speed is a property of the myosin isoform expressed in a muscle, we investigated how sequence changes in a range of muscle myosin II isoforms enable this slower rate of muscle contraction. We considered 798 sequences from 13 mammalian myosin II isoforms to identify any adaptation to increasing body mass. We identified a correlation between body mass and sequence divergence for the motor domain of the four major adult myosin II isoforms (β/Type I, IIa, IIb, IIx), suggesting that these isoforms have adapted to increasing body mass. In contrast, the non-muscle and developmental isoforms show no correlation of sequence divergence with body mass. Analysis of the motor domain sequence of β-myosin (predominant myosin in Type-I/slow and cardiac muscle) from 67 mammals from two distinct evolutionary clades identified 16 sites, out of 800, associated with body mass ($p_{adj}<0.05$) but not with the clade ($p_{adj}>0.05$). Both clades change the same small set of amino acids, in the same order from small to large mammals, suggesting a limited number of ways in which contraction velocity can be successfully manipulated. To test this relationship, the nine sites that differ between human and rat were mutated in the human β-myosin to match the rat sequence. Biochemical analysis revealed that the rat-human β-myosin chimera functioned like the native rat myosin with a two-fold increase in both motility and in the rate of ADP release from the actin.myosin cross-bridge (the step that limits contraction velocity). Thus these sequence changes indicate adaptation of β-myosin as species mass increased to enable a reduced contraction velocity and heart rate.

*Corresponding author, e-mail: N.Wass@kent.ac.uk

Deciphering cancer genomes

Mona Singh

Lewis-Sigler Institute for Integrative Genomics, Department of Computer Science, Princeton University, Princeton, USA

Abstract

Large-scale cancer genome sequencing consortia have provided a huge influx of somatic mutation data across large cohorts of patients. Understanding how these observed genetic alterations give rise to specific cancer phenotypes is a major aim of cancer genomics. This is challenging because numerous somatic mutations occur in each cancer genome, but only a subset are cancer-relevant; further, there is a high degree of mutational heterogeneity across individuals. Fortunately, the large and diverse biological datasets collected over the past few decades—including genome sequences across organisms and healthy individuals, protein structural data and interaction networks—provide a rich context within which to interpret cancer mutational data. In this talk, I will overview integrative computational methods my group has developed to interpret cancer mutational data, with an emphasis on identifying interactions perturbed in cancers.

*Corresponding author, e-mail: msingh@cs.princeton.edu

Meta-analysis of circulating cell-free DNA's role in the prognosis of pancreatic cancer

NATAŠA MILIĆ
Faculty of Medicine, University of Belgrade, Serbia

Abstract:
INTRODUCTION: Analyzing cell free DNA (cfDNA) for genetic abnormalities is a new prom-ising approach for diagnosis and prognosis of pancreatic cancer patients. Insights into the mo-lecular characteristics of pancreatic cancer may provide valuable information, leading to its ear-lier detection and development of targeted therapies. MATERIAL AND METHODS: We con-ducted a systematic review and meta-analysis of studies that reported cfDNA in pancreatic ductal adenocarcinoma (PDAC). Studies were considered eligible if they included patients with PDAC, studies having blood tests for cfDNA/ctDNA, and studies analyzing the prognostic value of cfDNA/ctDNA for patients' survival. Studies published before October 22, 2020 were identified through PubMED, EMBASE, Web of Science and Cochrane Library databases. The assessed out-comes were overall (OS) and progression-free survival (PFS) expressed as the log Hazard Ratio (HR) and standard error (SE). The summary HR effect size was estimated by pooling individual trials results using the Review Manager, version 5.3, Cochrane Collaboration. Heterogeneity was assessed using the Cochran Q test and I2 statistic. RESULTS: 48 studies were included in qualitative review, while 44 were assessed in quantitative synthesis, with total number of pa-tients included 3524. An overall negative impact of cfDNA and KRAS mutations on OS and PFS in PDAC (HR=2.42, 95%CI 1.95–2.99 and HR=2.46, 95%CI: 2.01–3.00, respectively) were found. Subgroup analysis of locally advanced and metastatic disease presented similar results (HR=2.51, 95%CI: 1.90-3.31). In studies assessing pre-treatment presence of KRAS, there was a moderate to high degree of heterogeneity (I2=87% and I2=48%, for OS and PFS, respectively), which was remarkably decreased in analysis of studies measuring post-treatment KRAS (I2=24% and I2=0%, for OS and PFS, respectively). Patients who were KRAS positive before but KRAS negative after treatment had a better prognosis than persistently KRAS positive patients (HR=5.30, 95%CI: 1.02–27.63) CONCLUSION: Assessing KRAS mutation by liquid biopsy can be considered as an additional tool for estimating disease course and outcome in PDAC patients.

Keywords:
cell free DNA; pancreatic ductal adenocarcinoma; survival; meta-analysis

*Corresponding author, e-mail: silly_stat@yahoo.com

Network data fusion and topological analysis identifies the neighbours of viral targets and differentially expressed genes in Covid-19 as drug target candidates

Natasa Przulj[1, 2, 3]

[1] Catalan Institution for Research and Advanced Studies (ICREA), Spain
[2] Barcelona Supercomputing Center, Spain
[3] University College London, UK

Abstract:
The COVID-19 pandemic is raging. It revealed the importance of rapid scientific advancement towards understanding and treating new diseases. To address this challenge, we build onto our previous methods for extracting new biomedical knowledge from the wiring patterns of systems-level, heterogeneous biomedical networks. These methods are needed due to the flood of molecular and clinical data, measuring interactions between various bio-molecules in and around a cell that form large, complex systems. These systems-level network data provide heterogeneous, but complementary information about cells, tissues and diseases. The challenge is how to mine them collectively to answer fundamental biological and medical questions. This is nontrivial, because of computational intractability of many underlying problems on networks (also called *graphs*), necessitating the development of approximate algorithms (heuristic methods) for finding approximate solutions.

We will give an overview of the lab's work. Also, we will focus on explaining how we adapt an explainable artificial intelligence algorithm for data fusion and utilize it on new omics data on viral-host interactions, human protein interactions, and drugs to better understand SARS-CoV-2 infection mechanisms and predict new drug-target interactions for COVID-19. We discover that in the human interactome, the human proteins targeted by SARS-CoV-2 proteins and the genes that are differentially expressed after the infection have common neighbors central in the interactome that may be key to the disease mechanisms. We uncover 185 new drug-target interactions targeting 49 of these key genes and suggest re-purposing of 149 FDA-approved drugs, including drugs targeting VEGF and nitric oxide signaling, whose pathways coincide with the observed COVID-19 symptoms. Our integrative methodology is universal and can enable insight into this and other serious diseases, as well as personalize treatment.

*Corresponding author, e-mail: natasha@bsc.es

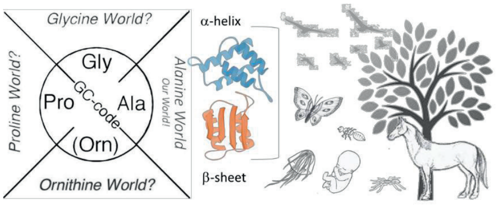
Structure and Evolution of the Genetic Code from the Perspective of a Skeptical Experimental Biochemist

Nediljko Budisa[1,2]

[1]University of Manitoba, Department of Chemistry & Microbiology, Chair of Chemical Synthetic Biology 144 Dysart Rd, R3T 2N2 Winnipeg, MB, Canada,
[2]Berlin Institute of Technology/TU Berlin, Department of Chemistry, Biocatalysis Group Müller-Breslau-Straße 10, D-10623 Berlin, Germany.

Abstract
Life on earth is a unity owing to the existence of the universal genetic code, i.e. as the genetic code for all organisms is basically the same - all living things use the same "genetic language". This biological framework allows the translation of the genetic message written in DNA into life-sustaining proteins build with 20 canonical amino acids. Why 'only' 20? According to Woese's early suggestion, primitive cells or protocells began with a completely random, highly ambiguous set of codon assignments to amino acids with very inaccurate translation. The process of expanding the amino acid repertoire ended with the "frozen accident". Therefore, the answer to the question why only 20 amino acids are the standard repertoire of the universal genetic code is an evolutionary one.

Recently, we proposed Alanine World model that explains the choice of amino acids monomers in the genetic code repertoire. The core of our considerations is the selection of secondary structural elements for the construction of the protein scaffolds and the subsequent life body plans (forms, morphology). This also determines the choice of amino acid monomers. Dominant secondary structures in life as we know it are α-helices and β-sheets, which are mainly made up of alanine derivatives (from a chemical point of view; that is why we coined the term "Alanine World"). In this model, the selection of monomers (amino acids) for such secondary structural elements is rather limited by for α-helix or β- sheet propensities. The choice of alanine also results from the existing universal core metabolism of the earth, which constantly supplies basic metabolic intermediates and suitable building blocks. The Alanine World model, together with the RNA world hypothesis and the Co-evolution theory, offers a plausible scenario for the chemical etiology of the amino acid repertoire in the genetic code. The earlier GC code (in the RNA world) with only 4 amino acid building blocks (Gly, Ala, Pro and Arg equivalents such as Ornithine) has chosen among these few options the Alanine World. The fundamental question now is whether we can revisit these options from the evolutionary past and experimentally create parallel biological Proline-, Glycine- or Ornithine-Worlds based on fundamentally different chemistry, metabolism, energetics of life with dramatically altered genetic codes. Systems bioengineering (Synthetic Biology and Xenobiology) of the 21st century will answer this question for us.



Kubyshkin, V., Budisa, N: The Alanine World Model for the Development of the Amino Acid Repertoire in Protein Biosynthesis. *Int. J. Mol. Sci.* **2019**, *20(21)*, e5507.

*Corresponding author, e-mail: nediljko.budisa@umanitoba.ca

New age for alignment-free methods for sequence analyses

Vladimir Perovic, Branislava Gemovic and Nevena Veljkovic*

Laboratory for bioinformatics and computational chemistry, VINCA Institute of Nuclear Sciences, University of Belgrade, National Institute of the Republic of Serbia

Abstract:
Progress in a wide range of fields ranging from population genetics to precision medicine may be attributed to availability of big biological data. Alignment-free sequence comparison is the methodology of choice in data-intensive applications given that it is significantly faster and requires less resources compared to traditional sequence comparison based on pairwise or multiple sequence alignment.

The symbiosis of alignment-free methods with machine learning is a paradigm of new age in bioinformatics, as it ensures the much needed boost to quicken the complex predictions on large datasets, particularly of molecules with low sequence identity.

In this talk, I will present two stories in which I will describe approaches to predict functional consequences of gene variants and imperfect tandem repeats in protein sequences.

*Corresponding author, e-mail:

Exploring amyloidogenicity of peptides from ribosomal protein S1 to develop novel AMPs

Oxana V. Galzitskaya [1,2]

[1]Laboratory of Bioinformatics and Proteomics, Institute of Protein Research, Russian Academy of Sciences, Pushchino, Moscow Region, Russia
[2]Laboratory of the structure and function of muscle proteins, Institute of Theoretical and Experimental Biophysics, Russian Academy of Sciences, Pushchino, Moscow Region, Russia

Abstract
Antimicrobial peptides (AMPs) and similar compounds are potential candidates for combating antibiotic-resistant bacteria. The hypothesis of directed coaggregation of a target protein and an amyloidogenic peptide acting as an antimicrobial peptide was successfully tested for peptides synthesized on the basis of ribosomal protein S1 in a bacterial culture of T. thermophilus. The coaggregation of the target protein and the amyloidogenic peptide was also tested for the pathogenic P. auregenosa. Almost all peptides that we selected as AMPs prone to aggregation and formation of fibrils based on the amino acid sequence of ribosomal protein S1 formed amyloid fibrils. We demonstrated that amyloidogenic peptides are not only toxic to its host target cells but also some of them possess antimicrobial activity. This direction is promising for the opening of new AMPs. Controlling the aggregation of vital bacterial proteins can become one of the new directions of research and form the basis for the search and development of targeted antibacterial drugs.

*Corresponding author, e-mail: ogalzit@vega.protres.ru

Extending the small-molecule similarity principle to all levels of biology

Patrick Aloy

Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain

Abstract

Small molecules are usually compared by their chemical structure, but there is no unified analytic framework for representing and comparing their biological activity. In this talk, I will present the Chemical Checker (CC), which provides processed, harmonized and integrated bioactivity data on ~1M small molecules. The CC divides data into five levels of increasing complexity, from the chemical properties of compounds to their clinical outcomes. In between, it includes targets, off-targets, networks and cell-level information, such as omics data, growth inhibition and morphology. We show how CC signatures can aid drug discovery tasks, including target identification and library characterization. We also demonstrate the discovery of compounds that reverse and mimic biological signatures of disease models and genetic perturbations in cases that could not be addressed using chemical information alone. Overall, the CC signatures facilitate the conversion of bioactivity data to a format that is readily amenable to machine learning methods.

*Corresponding author, e-mail: paloy@irbbarcelona.org

Biomolecular condensation: macromolecular assembly in the cell by liquid–liquid phase separation and beyond

Peter Tompa[1,2,3]

[1] *VIB, VIB-VUB Center for Structural Biology, Brussels, Belgium*
[2] *Vrije Universiteit Brussel (VUB), Brussels, Belgium*
[3] *Institute of Enzymology, Research Centre for Natural Sciences, Budapest, Hungary*

Abstract

Biomolecular condensation is a process whereby a large number of macromolecules (proteins and RNA) form a non-stoichiometric, functional assembly. The dominant mechanism of such association is liquid-liquid phase separation (LLPS), which leads to the formation of membraneless organelles (MLOs), such as the nucleolus and stress granules, in the cell. The proteins involved often have a high proportion of intrinsic structural disorder (ID), which drive LLPS by transient, multivalent interactions.

In this presentation I outline three specific, thus far little appreciated aspects of physiological phase separation. First, condensates represent a special emergent functional state of ID proteins (IDPs), which is manifested not at the level of individual proteins. Second, the formation of condensates is highly regulated at many levels, and "phase separating" proteins have very different contributions to LLPS, which translates to their different physicochemical properties. Finally, I will show that there are numerous cellular examples, which are highly dynamic, yet not liquid, rather assume a "pleiomorphic" structural state, which may represent the most important organizational principle in the cell.

*Corresponding author, e-mail: peter.tompa@vib-vub.be

Integrating generative physical and deep learning approaches to navigate the structure-function-sequence triangle

Richard Bonneau

Center for Genomics and Systems Biology, New York University, USA

Abstract

I will discuss recent work to integrate machine learning methods into overarching methods to predict and design protein function. I will demonstrate that these methods also lead to new, more scalable, methods to align, classify and organize large sequence databases. First we will discuss methods for using language models to extract features from sequences that allow for massive scaling and improvement of alignment based computational methods. This will culminate in our method for end to end differentiable alignment, deepBlast. I will then discuss the work of the eminent scholar Vladimir Gligorijevic on using structure to build better models of protein function-sequence-structure relationships. This culminates in a method called DeepFRI. This method learns models of protein structure-function relationships that also localize function to punctuate (and correct) regions on the protein structure. Co authors on this work also include, but are not limited to: Julia Koehler, Vikram Mulligan, Dan Berenberg, Meet Barot, Jamie Morton and Kyunghyun Cho.

*Corresponding author, e-mail: bonneau@nyu.edu

Utilizing multi-omics and medical records to understand cancer

Ron Shamir

The Blavatnik School of Computer Science, Tel Aviv University, Israel

Abstract
Today's large biological datasets open novel opportunities in basic science and medicine. While inquiry of each dataset separately often provides insights, integrative analysis may reveal more holistic, systems-level findings. We demonstrate the power of integrated analysis in cancer on three levels: (1) in joint analysis of multiple omics for the same cancer; (2) in identifying and ranking driver genes in an individual's tumor based on expression and mutation profiles; and (3) in predicting and individual's future risk of developing cancer based on medical records of routine periodical checkups. In all cases, we develop novel methods and observe a clear advantage of the integration.

*Corresponding author, e-mail: rshamir@tau.ac.il

Genome as a program

Sergei Kozyrev[1]

[1] *Steklov Mathematical Institute, Gubkina 8, Moscow, Russia*

Abstract

We consider a model of genome as a program which operates by parallel application of genes. Genes are subject to gene regulation. We discuss genome as a functional program written in Haskell-like language: parallel functioning of genes is described by recursive application of list of functions (genes) as applicative list functor and gene regulation is described by monadic computations. Darwinian evolution in this picture is described as learning for functional programs, in particular learning model for functional programs is formulated.

Keywords:

bioinformatics, computer science, machine learning

*Corresponding author, e-mail: kozyrev@mi-ras.ru

Recent Advances in Large-scale Protein Function Prediction

Shanfeng Zhu

Institute for Science and Technology for Brain-Inspired Intelligence, and the Shanghai Institute of Artificial Intelligence Algorithms at Fudan University, China

Abstract
Proteins are building blocks of life, playing many crucial roles within organisms, such as catalysing chemical reactions, coordinating signal pathway and providing structural support to cells. Automated function prediction (AFP) of proteins is thus of great significance in biology. AFP can be regarded as a problem of the large-scale multi-label classification where a protein can be associated with multiple gene ontology terms as its labels. To boost the development of effective and efficient AFP, Critical Assessment of Functional Annotation (CAFA) has been held four times to date: CAFA1 in 2010–2011, CAFA2 in 2013–2014, CAFA3 in 2016–2017 and CAFA4 in 2019–2020 (under evaluation). In this talk, I will introduce the state-of-the-art methods in large-scale AFP, as well as our recent progress in this topic, such as GOLabeler, NetGO and DeepGraphGO.

*Corresponding author, e-mail: zhusf@fudan.edu.cn

The inquiry of the hidden information of cis-regulatory elements

Silvano Piazza

International Centre for Genetic Engineering and Biotechnology, University of Trento, Trento, Italy

Abstract:
The FANTOM5 project generated a unique resource, the first single molecule sequencing-based expression atlas in mammalian systems. Cap analysis of gene expression (CAGE) was used to measure transcription start sites (TSS) and promoter usage across a collection of over 1000 samples in human and mouse thereby identifying and measuring levels of the majority of coding and non-coding transcripts in the mammalian genome. In this work, we analyzed the FANTOM5 human dataset using ScanAll, a newly developed software here described, to ab initio predict the presence of enriched elements in the FANTOM5 promoters. First, we identified motifs enriched in a subset of genomic regions, possibly corresponding to Transcription Factor Binding Sites (TFBS); then we pinpointed the existence of structured regulatory modules, i.e. groups of enriched motifs co-occurring in co-expressed regions within a flexible distance. We associated differences in modularity with changes occurring in the nucleotide content of the examined regions, sub-classifying FANTOM promoters in eight categories associated with different expression levels and tissue specificity, leading to new insights in the constitutive and tissue-specific genes regulation.

*Corresponding author, e-mail: Silvano.Piazza@icgeb.org

Network science reveals a protein's role in aging and a person's risk of mental health problems

Tijana Milenkovic

Head of the Complex Networks (CoNe) Lab, University of Notre Dame, USA

Abstract

Networks (or graphs) are powerful models of complex systems in various domains, from biological cells to societies to the Internet. How to efficiently study these data, especially with increasing availability of dynamic (temporal) real-world networks? This talk will discuss our state-of-the-art computational approaches for network analysis, including those for studying dynamic networks, some of which are based on graphlets (subgraphs, Lego-like building blocks of complex networks). Also, this talk will demonstrate usefulness of our (dynamic) network analysis approaches in two tasks: studying the role of a protein in the aging process based on its position in a (dynamic) molecular network, and analyzing an individual's mental health based on their position in a (dynamic) social network, as follows.

First, incidence of many complex diseases, such as cancer, Alzheimer's disease, and even COVID-19 increases with age. Understanding the molecular mechanisms behind the aging process, including identification of human genes (i.e., their protein products) implicated in aging, is important for treating such aging-related diseases. However, wet lab experimental analyses of human aging are hard due to long human life span and ethical constraints. Computational identification (i.e., prediction) of aging-related genes via machine learning from human -omics data can fill in this gap. In this context, we integrated aging-specific gene expression data with context-unspecific protein-protein interaction (PPI) network data to infer a dynamic aging-specific PPI subnetwork. Then, we developed a machine learning model that when applied to the dynamic subnetwork can analyze how genes' PPIs change with age. So, our predictive model could guide the discovery of novel aging-related gene candidates for future wet lab validation.

Second, mental disorders such as depression and anxiety are public health issues. Early interventions can significantly reduce risk of developing mental disorders. Yet, most people do not seek treatments due to a lack of awareness of their disorders. A way to raise awareness is to develop computational approaches for predicting whether and when an individual will become at risk of a mental disorder. Innovative technologies such as wearable sensors can provide a wealth of data relevant to mental health. In this context, we leveraged rich longitudinal data from the recent NetHealth study containing individuals' social interaction data collected via smartphones, health-related behavioral data (physical activity and sleep duration) collected via Fitbit devices, and a variety of individuals' trait data (including mental health) collected via surveys. We modeled the NetHealth data as a dynamic network and developed a machine learning model for predicting one's likelihood of being depressed or anxious based on how their position in the dynamic network changes over time.

*Corresponding author, e-mail: tmilenko@nd.edu

cAMP mediated genes profiles underscore drastic reduction of dopamine intake along with opioids in the dorsal striatum of fighting deprived aggressive mice

Vladimir Babenko*, Dmitry Smagin, Irina Kovalenko, Anna Galyamina, and Natalia Kudryavtseva

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Abstract
In accordance with our previous study (Babenko et al., 2019; PMID: 32216748), we assessed the cAMP mediated genes network in the samples of the dorsal striatum in male mice with aggression experience during 20 day agonistic interactions before (A20) and after period of aggression deprivation (AD).

We found that dopamine uptake and, specifically, consequent opioid synthesis highly presented in A20 species in dorsal striatum neurons were totally (significantly) curbed in AD mice. Along with that, AD group demonstrated higher aggression rate in fighting sessions than A20 group. It implies aggression mediated opioid withdrawal (toxic) syndrome in AD species upon fighting deprivation outlined earlier for this model (Kudryavtseva et al., 2014; PMID: 25340443).

Herein, as we kept on modifying the genes system of cAMP response network in striatum medium spiny neurons (MSN) in (Babenko et al., 2019; PMID: 32216748), we should underscore the role of *Grin1* (NR1) alternative splicing specifics. We found its consistent isoform – specific preference in certain stages and MSN types. In particular, it manifests specific glutamate *Grin1* isoform comprising exon 4, along with *Grin2a*, *Grin2c* NMDA subunits upon Medium Spiny *Drd1*-neurons activation of firing stage mediated by PKA cascade, while *Drd2* MSNs use only exon 4-lacking *Grin1* isoforms and *Grin2b* subunit upon firing stage in NMDA receptors.

Additionally, we elaborated on the proteome diversity in the dorsal striatum of aggressively experienced animals, outlining the peculiarities and complexities of the neuronal genes involved. We found confirmation of aggression withdrawal status in animals based on the key addiction mediated gene network of cAMP response in this brain region.

The study was funded by Russian Science Foundation project 19-15-00026.

Keywords:
chronic social conflict model, RNA-Seq, dopamine, glutamate, dorsal striatum, cAMP network.

*Corresponding author, e-mail: bob@bionet.nsc.ru

Supervised Machine Learning Techniques for Single Cell RNAseq:New Opportunities for Clinical Medicine

Vladimir Brusic

Li Dak Sum Chair Professor in Computer Science
School of Computer Science, University of Nottingham Ningbo China

Abstract
Single cell transcriptomics (SCT) detects gene expression from individual cells. Bulk sequencing from mixed samples provides only average gene expression across all cells in the sample. SCT provides information about the heterogeneity of gene expression within cell types and subtypes and various healthy or disease states. SCT data sets are presented as sparse matrices with more than 30,000 genes in matrix rows, and up to a million cells in columns. These number and the size of these data sets are growing at exponential rate both in the number of cells per matrix, and the number of data sedts. The analysis of gene expression from single cells is essential for understanding the cellular and molecular basis of biological and pathological processes. We cleaned and standardized more than 2000 SCT data sets and developed a system that classifies peripheral blood cells by artificial neural networks. The accuracy of classification of peripheral blood cells reached 95% in 5-class classification. For tissue of origin classification, we achived accuracy of 80% for 10-class classification and 98% in 7-class classification (when 5 tissues were combined in 2 superclasses). We also demonstrated the ability to distinguish chronic lymphocytic leukemia samples before treatment and at various time points after the start of treatment. The differences between healthy and disease states are reflected in differential gene expression across multiple cell types and subtypes. These differences can be used for diagnosis, prognosis, and therapy selection in cancer, infectious disease, autoimmunity, and other pathological states.

*Corresponding author, e-mail: vladimir.brusic@nottingham.edu.cn

Unusual Biophysics and Strange Biology of Intrinsic Disorder

Vladimir N. Uversky

Department of Molecular Medicine, University of South Florida, Tampa, Florida 33612, USA

Abstract

Intrinsically disordered proteins (IDPs) lack stable tertiary and/or secondary structure under physiological conditions *in vitro*, often resembling 'protein clouds'. Computational studies revealed that IDPs are highly abundant in nature, as ~25-30% of eukaryotic proteins are mostly disordered, and >50% of eukaryotic proteins and > 70% of signaling proteins have long disordered regions. The functional repertoire of IDPs is complementary to that of ordered proteins, with IDPs being commonly involved in regulation, signaling and control pathways, where binding to multiple partners and high-specificity/low-affinity interactions play a crucial role. It is suggested that functions of IDPs may arise from the specific disorder form, from interconversion of disordered forms, or from transitions between disordered and ordered conformations. The choice between these conformations is determined by the peculiarities of the protein environment, and many IDPs possess an exceptional ability to be highly responsive to change in their environment and to fold in a template-dependent manner. All this requires a close attention to the odd biophysics of IDPs. In this talk, some key biophysical features of IDPs will be covered. In addition to the peculiar sequence characteristics these unusual biophysical features include sequential, structural, and spatiotemporal heterogeneity of IDPs; their rough and relatively flat energy landscapes; their ability to undergo both induced folding and induced unfolding; the ability to interact specifically with structurally unrelated partners; the ability to gain different structures at binding to different partners; and the ability to keep essential amount of disorder even in the bound form. IDPs are also characterized by the "turned-out" response to the changes in their environment. It is proposed that the heterogeneous spatiotemporal structure of IDPs/IDPRs can be described as a set of foldons, inducible foldons, semi-foldons and non-foldons. They may lose their function when folded, and activation of some IDPs is associated with the awaking of the dormant disorder. IDPs are tightly controlled in the norm by various genetic and non-genetic mechanisms. Alteration in regulation of this disordered regulators are often detrimental to a cell and many IDPs are associated with a variety of human diseases such as cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases, diabetes and others. Therefore, there is an intriguing interconnection between intrinsic disorder, cell signaling and human diseases. Pathogenic IDPs, such as α-synuclein, tau protein, p53, BRCA1 and many other disease-associated hub proteins represent attractive targets for drugs modulating protein-protein interactions. Several strategies have been elaborated for elucidating the mechanisms of blocking of the intrinsic disorder-based protein-protein interactions.

*Corresponding author, e-mail: vuversky@usf.edu

Reconstruction of Gene Networks Associated with Complex Disorders on Example of Parkinson Disease

Yuriy L. Orlov[1,2*], Ayya G. Galieva[2], Anton N. Luzin[2], Anastasia A. Anashkina[1,3]

[1] The Digital Health Institute, I.M.Sechenov First Moscow State Medical University (Sechenov University), Trubetskaya 8-2, 119991 Moscow, Russia
[2] Novosibirsk State University, Pirogova, 1, 630090 Novosibirsk, Russia
[3] Engelhardt Institute of Molecular Biology RAS, Vavilova, 32, 199991 Moscow, Russia

Abstract

Reconstruction of gene networks underlying molecular mechanisms of gene expression regulation in complex diseases may help find new targets for therapy. Analysis of common genes in the networks for different diseases allows to find connections between different diseases, describe their statistical properties. The accumulation of genetic data in the field of Parkinson's disease research is currently progressing a lot from identifying risk factors to confident prediction of the disease occurrence.

To find new gene-targets for diagnostics and therapy we have to reconstruct gene network of the disease, to cluster genes in the network, to reveal key (hub) genes with largest number of interactions in the network. Using on-line bioinformatics tools OMIM, PANTHER, g:Profiler, GeneMANIA, and STRING-DB, we analyzed the current array of data related to Parkinson's disease, and other complex disorders such as schizophrenia. We calculated the categories of gene ontologies for a large list of genes, visualized them, and built gene networks containing the identified key objects and their relationships.

We reconstructed gene networks for complex mental disorders and diseases and estimated statistical parameters of the network (connectivity, clustering) as well as annotated hub genes. However, translating the results into biological understanding is still a promising major challenge. The analysis of the genes associated with the disease, the assessment of their place in the gene network (connectivity) allows us to evaluate them as target genes for medicinal effects.

Keywords:

bioinformatics, Parkinson's disease, reconstruction of gene networks, gene ontology

*Corresponding author, e-mail: orlov@d-health.institute

Influence of medical domain knowledge on deep learning for early diagnostics of Alzheimer's disease

Zoran Obradovic

Data Analytics and Biomedical Informatics Center, Computer and Information Sciences Department, Statistics Department, Temple University

Abstract:
An early diagnosis of Alzheimer's Disease (AD), the sixth leading causes of death for adults, provides a better chance of benefiting from treatment. The objective of our study was to determine if an accurate and stable AD risk scoring method could be achieved without relying on costly imaging-based diagnostics. The proposed approach extracts information from the repository of about 40 million heterogeneous ambulatory medical records that include patients' conditions, procedures, measurements, and drug domains and applies clinical domain knowledge in data preprocessing and positive cohort selection for training a Long-Short-Term Memory (LSTM) Recurrent Neural Network (RNN) deep learning model to predict will the patient develop AD. The LSTM RNN method that used integrated data relevant to AD performed significantly better then learning from the cohorts selected naïvely achieving clinically relevant out-of sample score (AURPC above 0.99).

Reported results are based on the following paper: Ljubic, B., Roychoudhury, S., Cao, X., Pavlovski, M., Obradovic, S., Nair, R., Glass, L, Obradovic, Z. "Influence of Medical Domain Knowledge on Deep Learning for Alzheimer's Disease Prediction," *Computer Methods and Programs in Biomedicine*, v. 197, Dec. 2020, 205765.

*Corresponding author, e-mail:

INVITED SPEAKERS:

Structure Predictions Transform Protein Family Classification

Alex Bateman

Protein sequence resources, EMBL-EBI, UK

Abstract

Structural prediction models have come of age and are beginning to revolutionise molecular biology. In the recent CASP competition AlphaFold 2 showed accuracies close in many cases to crystal structures. Other methods such as trRosetta and RaptorX still give excellent models that are adequate for many applications. In this talk I will discuss how in collaboration with David Baker's group we released a large collection (>6,300) of structural models for Pfam families. We have begun to dig into this treasure trove to refine, define and classify protein domain families.

Pfam is a collection of over 19,000 protein families with multiple sequence alignments and profile-HMMs. Pfam is widely used to annotate genomes and metagenomes. Ideally Pfam families would correspond to structural domains or repeats found in protein structures. However, often the Pfam family was built before a structure was known. They may be truncated single domains or contain multiple domains. We find that protein structural models can be used to split large multidomain families. Structural models based on Pfam alignments have been less useful to correct truncated domains and these will require the creation of structural models of full length proteins. The structural models have also been instrumental in identifying which superfamilies many Pfam's should belong to. Thus the difference between protein sequence and protein structure classification is becoming smaller and may be unified within the coming few years.

*Corresponding author, e-mail: agb@ebi.ac.uk

De Novo Transcriptome Sequencing of *Ramonda serbica*: Identification of Late Embryogenesis Abundant Proteins

Ana Pantelić[1]¸ Strahinja Stevanović[2], Nataša Kilibarda[3], Marija Vidović [2*]

[1] University of Belgrade, Faculty of Chemistry, Studentski trg 12-16, 11000 Belgrade, Serbia
[2] Institute of Molecular Genetics and Genetic Engineering, Laboratory for Plant Molecular Biology, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia
[3] Singidunum University, Danijelova 32, 11000 Belgrade, Serbia

Abstract

An extreme loss of cellular water or desiccation (5-10% of relative water content) leads to protein denaturation, aggregation and degradation, and affects the fluidity of membrane lipids resulting in loss of membrane integrity [1]. The essential constituents of vegetative desiccation tolerance in so-called resurrection plants are late embryogenesis abundant proteins (LEAPs). This heterogeneous group of anhydrobiosis-related intrinsically disordered proteins forms mostly random conformation when fully hydrated, turning into compact α-helices during desiccation [2]. Based on *in vitro* studies, LEAPs can be involved in water binding, ion sequestration, stabilization of both membrane and enzymes during freezing or drying, while by forming intracellular proteinaceous condensates they increase structural integrity and intracellular viscosity of cells during desiccation.

Here, we identify 164 members of LEA gene family in endemic and relict resurrection species *Ramonda serbica* by integrating previously done *de novo* transcriptome and homologues protein motifs. Identified LEAPs were classification into six groups according to Protein family (PFAM) database and the most populated group was LEA4 containing 47% of total identified LEAPs. By using four secondary structure predictors, we showed that this group exhibited a high propensity to form amphipathic α-helices (81% of total sequence length is predicted to form α-helical structure). This implies that charged residues might be exposed to the solvent, while hydrophobic amino acids might interact with lipid bilayers or with other target proteins in the cell. In addition, as predicted by several bioinformatics tools, more than 70% of identified LEAPs were found to be highly disordered. Structural characterization of LEAPs is a key to understand their function and regulation of their intrinsic structural disorder-to-order transition during desiccation. These findings will promote transformative advancements in various fields, such as the development of new strategies in neurodegenerative disorders, cell preservation technology and the improvement of crop drought tolerance.

Keywords:

desiccation tolerance, intrinsically disordered proteins, liquid-liquid phase separation, resurrection plants, secondary structure prediction, water stress

Acknowledgements:

This work was funded by the Science Fund of the Republic of Serbia (PROMIS project LEAPSyn-SCI, grant number 6039663) and by the Ministry of Education, Science and Technological Development, the Republic of Serbia (Contract No. 451-03-9/2021-14/ 200042).

*Corresponding author, e-mail: mvidovic@imgge.bg.ac.rs

Inflammatory bowel disease prediction based on metagenomics data

Andrea Mihajlović[1*], Katarina Mladenović[2] , Tatjana Lončar-Turukalo[2], Sanja Brdar[1]

[1] BioSense Institute, University of Novi Sad, dr Zorana Đinđića 16,
21000 Novi Sad, Serbia
[2] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6,
21000 Novi Sad, Serbia

Abstract
Inflammatory bowel disease (IBD) is a genetic disease manifested under certain external influences. Traditional disease models aim to find a single pathogen which causes disease. However, many studies have revealed that no single pathogen causes IBD. Machine learning algorithms applied on microbiome data have huge potential in uncovering patterns and aiding diagnosis of diseases including IBD. In this study we investigate microbiome variations in stool samples, in an attempt to evaluate performance of classification algorithms in identifying IBD state.

Dataset used in this study (available from Integrative Human Microbiome Project) contains 429 samples from 27 healthy subjects, and 1209 samples from 103 IBD subjects. Microbes are grouped into 1479 operational taxonomic units (OTUs) and in this form used in our analysis. In preprocessing, data was log2-transformed. From many investigated algorithms, a random forest (RF) classifier was selected for detailed evaluation in a binary (IBD, nonIBD) classification task. The class imbalance was approached using balanced RF (BRF) which under-samples the majority class in a bootstrap process. Parameter searching was conducted using a cross-validation approach. Initial set of parameters was created at random, further evaluated through grid search and fine tuned in the vicinity of best performing parameters. Dimensionality was reduced by searching for the smallest feature subset which preserves the performance. Experiments included hand-picked taxa and/or selected *k* best scoring features. Training was performed for each model in 100 iterations with 10-fold cross-validation, which ensured comprehensive evaluation. Upon sample-wise binary classification, subjects were labelled as IBD based on average decision probability of their samples, by varying different thresholds. Change in classification performance as a function of the employed threshold was noted.

Best model comprised 150 trees with a maximum depth of 15, using entropy for node splitting. With the average *F1* score of 94% our study confirms the strong connection of IBD and gastrointestinal microbiome. Retraining model using 100 most important features showed minor decrease in *F1* score of 1% and exclusion of all strains and organisms other than bacteria showed no decrease. Further research efforts should focus on gathering more data and improved model explainability in predicting IBD state.

Keywords:
microbiome, OTU table, machine learning, features selection

*Corresponding author, e-mail: andrea.mihajlovic@biosense.rs

Understanding Infection Progression under Strong Control Measures through Universal COVID-19 Growth Signatures

Magdalena Djordjevic[1], Marko Djordjevic[2], Bojana Ilic (Blagojevic)[1*], Stefan Stojku[1], and Igor Salom[1]

[1] Institute of Physics Belgrade, Pregrevica 118, 11080 Belgrade, Serbia
[2]Quantitative Biology Group, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia

Abstract
While interventions such as quarantine or vaccination have been extensively studied within compartmental models in epidemiology, the effects of social distancing are poorly known. Even when considered, they were studied only numerically. Through joint analytical and numerical analysis, we developed a novel framework, which incorporates social distancing measures. The framework reproduces the main infection progression data (such as detected case counts, active cases and fatalities), capturing empirically observed COVID- 19 growth signatures of detected case counts, i.e., its three distinct dynamical regimes (exponential, superlinear and sublinear).

We utilized an approach well known to theoretical physics and more recently systems biology, where we look at common dynamical features, regardless of the differences in other factors. This approach provides generality of the applied framework, ensuring the applicability to a wide range of countries and other infectious diseases. The dynamical features and associated scaling laws are used as a powerful tool to pinpoint regions where analytical derivations are effective for i) imposing stringent restraints on parameter quantifying the effect of social distancing; ii) explaining the nearly constant value of the scaling exponent in the superlinear regime of detected counts; iii) understanding the relationship between the duration of this regime and strength of social distancing; iv) identifying changes in the reproduction number from outburst to extinguishing the infection. Additionally, we successfully applied this tool to infer key infection parameters. The main advantage of our analytically tractable model (compared to the state-of-the-art numerical simulations) is in its ability to qualitatively and quantitatively explain common dynamical features of a system, to yield a fundamental understanding of infection progression under strong control measures, and to provide highly constrained infection parameters inference.

Keywords:
systems biology, epidemiology, COVID-19, compartmental model, physics, social distancing measures, key infection parameters

*Corresponding author, e-mail: bojanab@ipb.ac.rs

Clustering as a Support Technique in Phenotyping and Genotyping

Dragana Bajić[1], Nataša Ž. Mišić[2*], Mirko Ostojić[2], and Nina Japundžić-Žigon[3]

[1] *Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia*
[2] *Research and Development institute Lola Ltd, Kneza Višeslava 70a,*
11000 Belgrade, Serbia
[3] *Faculty of Medicine, University of Belgrade, dr Subotića starijeg 8,*
11000 Belgrade, Serbia

Abstract

In the general rush of supervised machine learning implementations, clustering techniques remain almost unnoticed. However, these are the only methods suitable for multivariable analysis when available data are sparse. It is common for experiments with laboratory animals, the number of which is set by ethic directives to a minimum that ensures statitically reliable results.

This paper shows the support that clustering can offer to a biomedical experiment, aiming to define phenotype/s of delayed cardiomyopathy in adult male Wistar rats. Cardiomyopathy was induced by doxorubicin, an effective but toxic medication for treating malignancies. The "ground truth", established by echocardiography, revealed two groups of 9 rats each, one with preserved and the other with reduced ejection fraction. Two sets of features were associated with each rat: a) 18 physiological characteristics with pairwise correlation coefficient below 0.6 (echocardiography, blood pressure & heart rate variability, blood biochemistry); b) changes in the expression of 15 key cardiac genes, assessed by qPCR.

The clustering methods were hierarchical agglomerative clustering with a visual presentation of the cluster heat-map; K-mean ++ that minimizes the distance between the subjects and centroid; affinity propagation that forms a cluster around the most representative subject; Gaussian mixture model that uses an expectation-maximization algorithm, with probabilistic ("soft") cluster membership.

Algorithms are parametric and yielded 43 different results. The results were averaged using a simple framework of modified evidence accumulation clustering to form a co-existence matrix with elements grouped by hierarchical clustering to visualize the subject grouping. Visual grouping was also achieved by t-distributed stochastic neighbor embedding, a non-linear technique that maps each multidimensional point into the Cartesian coordinate system.

Both co-existence matrices and t-distributed embedding confirmed the existence of two major phenotypes. Implementation of the same techniques, however, yielded no corresponding genotype. A possible reason is that the genes responsible for phenotypic differences were not investigated in this paper. Although producing no coherent results if applied individually, the results confirmed that clustering techniques yield reliable results after the averaging by evidence accumulation. T-distributed embedding, with a proper choice of a distance metric, provides perfect grouping.

Keywords:
clustering, phenotype, genotype, doxorubicin.

*Corresponding author, e-mail: nmisic@rcub.bg.ac.rs

The emerging regulatory roles of noncoding RNAs in immune function of fish:microRNAs versus long noncoding RNAs

Haitham G. Abo-Al-Ela

Genetics and Biotechnology, Department of Aquaculture, Faculty of Fish Resources, Suez University, Suez, 43518, Egypt.

Abstract
The genome could be considered as raw data expressed in proteins and various types of noncoding RNAs (ncRNAs). However, a large portion of the genome is dedicated to ncRNAs,which in turn represent a considerable amount of the transcriptome. ncRNAs are modulated onlevels of type and amount whenever any physiological process occurs or as a response to external modulators. ncRNAs, typically forming complexes with other partners, are keymolecules that influence diverse cellular processes. Based on the knowledge of mammalian biology, ncRNAs are known to regulate and control diverse trafficking pathways and cellular activities. Long noncoding RNAs (lncRNAs) notably have diverse and more regulatory roles than microRNAs. Expanding these studies on fish has derived the same conclusion with relevance to other species, including invertebrates, explored the potentials to harness such typesof RNA to further understand the biology of such organisms, and opened gates for applying recent technologies, such as RNA interference and delivering micromolecules as microRNAs to living cells and possibly to target organs. These technologies should improve aquaculture productivity and fish health, as well as help understand fish biology.

Keywords:
fish biology; gene regulation; long noncoding RNAs; microRNAs;noncoding RNA; transcriptome

This manuscript with the same title has been published:
Abo-Al-Ela, H.G. The emerging regulatory roles of noncoding RNAs in immune function of fish: MicroRNAs versus long noncoding RNAs. Mol Genet Genomics (2021). https://doi.org/10.1007/s00438-021-01786-x

*Corresponding author, e-mail: haitham.aboalela@frc.suezuni.edu.eg

Entropy changes in worldwide collected HIV-1 subtypes A and B sequences

Ivan Skadric[1], Marina Siljic[1], Valentina Cirkovic[1], Maja Stanojevic[1]

[1]- University of Belgrade, Faculty of Medicine, Institute of Microbiology and Immunology, dr Subotica 8, 11000 Belgrade, Serbia.

Abstract

Evolutionary study of human immunodeficiency virus (HIV) has been ongoing for four decades, in pursuit of understanding the changes in its' genetics. Regarding selective pressure introduced by antiretroviral therapy (ART), it has been shown to depend on therapeutic regimen, but is also influenced by HIV-1 subtype. A common target for ART is viral protease, responsible for viral maturation and encoded by the pol gene. Protease inhibitors may show to be inefficient when certain mutations accumulate in the protease sequence in the pol gene. However, not all the changes on resistance-associated sites evolve in the same manner, while some became more frequent over time, others showed mild disturbance in frequency. To assess the extent of codon variation, Shannon entropy of nucleotide triplets was used as a measure of genetic diversity. Study dataset included HIV-1 pol sequences of subtypes A and B, from both naïve and therapy experienced patients, collected worldwide and obtained from the Stanford HIV drug resistance database. Sequences with large missing information caused by gaps or ambiguous nucleotides were removed. Codon-correct multiple sequence alignments were obtained by MAFFT service, resulting in subtype A and B alignment depths of 8119 and 68596 sequencies, respectively. To estimate entropy, a Python script was written that implements Shannon's approach on nucleotide sequences in a form of triplets, with those that do not match genetic code classified as a separate class (21st amino acid). After entropy rendering, sites that are related to drug resistance underwent hypergeometric testing for the following periods: before 1995, 1995 - 2009, and 2010 - 2017. The time periods were defined based on the major milestones of ART changes. Only those differences that breached Bonferroni corrected alpha level were considered statistically significant. Our results showed that, in general, frequencies of nucleotide triplets in HIV-1 subtypes A and B, differ between periods of different therapeutic implementation which is reflected in a slight entropy increase. In both subtypes, entropy reached the maximum (~3.6) at the 63rd triplet position. However, in HIV-1 subtype A, three drug resistance-related sites (48, 84, and 90) without statistically significant differences between the tested periods were observed.

Keywords:

Shannon's entropy, HIV, virus, data mining, drug resistance mutation.

*Corresponding author, e-mail: ivan.skadric@gmail.com

Virxicon: a lexicon of viral sequences

Jaroslaw Synak[1], Mateusz Kudla[1,2], Kaja Gutowska[1,3], Mirko Weber[2], Katrin Sophie Bohnsack[2], Piotr Lukasiak[1,3], Thomas Villmann[2], Jacek Blazewicz[1,3], Marta Szachniuk[1,3]

[1] *Institute of Computing Science and European Centre for Bioinformatics and Genomics, Poznan University of Technology, Poznan, 60-965, Poland.*
[2] *Saxon Institute for Computational Intelligence and Machine Learning, University of Applied Sciences Mittweida, Mittweida, 09648, Germany.*
[3] *Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, 61-704, Poland.*

Abstract

Viruses are ubiquitous in our world, they are present in almost every known environment. Biologists have been researching them for a very long time now, but there is still a lot to be discovered. This process involves processing a huge amount of information, especially during the pandemic, both gathering and exchanging data about viruses have become extremely important. Aforementioned data are scattered throughout different databases and it can be tedious for example to search for all viruses from a particular group. Our goal was to facilitate this process by building a comprehensive and easy to use database, which would present data from various viral datasets in a convenient form, and allow users to search by multiple different criteria (including Baltimore group). All data are fetched once a week from two well-known biological databases – NCBI and GENBANK, both eukaryotic and prokaryotic viruses are included. Virxicon allows users to access all the information using an intuitive web interface. It has both a mobile and a PC version freely available at http://virxicon.cs.put.poznan.pl/

Keywords:

database, RNA and DNA viruses, viral sequences, Baltimore classification, ICTV

*Corresponding author, e-mail: jaroslaw.synak@cs.put.poznan.pl

SB MultiCNV: novel method for copy number variations consensus calling

Jelisaveta Ilić[1], Vladimir Tomić[1*], Ana Popić[1*], Tea Kostić[1*], Nikola Škundrić[1*]

[1] *Seven Bridges, Omladinskih brigada 90, Serbia*

Abstract

Copy number variations (CNVs) represent types of structural variants involving alterations in the number of DNA segments, which can either be deleted or duplicated, leading to variable copy number in comparison to a normal genome. Numerous pathogenic CNVs have been connected with a variety of human diseases including diabetes, complex neurodegenerative and neuropsychiatric disorders, cancer and immune deficiency. A broad range of algorithms have been developed to detect CNVs from short-read sequencing data using read depth analysis. They all try to address some of the aforementioned problems, but their accuracy is not easy to measure nor compare. Therefore, defining optimal criteria for classifying CNV calls remains an open research problem in bioinformatics. In order to overcome these challenges, Seven Bridges has developed a novel method for CNV consensus calling, SB MultiCNV.

SB MultiCNV can process a multitude of output files from diverse CNV callers (Control FREEC, GATK4 CNV, PURPLE, Sclust, CNVKit, PureCN, CNVnator, Sequenza, FACETS) and works in two modes: *precise* mode which outputs high confidence regions where all callers are concordant and *majority* mode which outputs the regions for which the majority of the callers report matching copy number statuses.

SB MultiCNV is the only consensus caller which applies information about the ploidy of the sample to determine the status of each region. For each input file provided, a combination of regions was constructed and for each nucleotide position in the intersected regions, CNV call status was assigned to the CNV tool results and compared to the CNV call status of other inputs provided. For any region not covered by a tool-called CNV segment, but represented in other inputs, copy-number status was considered to be neutral.

SB MultiCNV has shown promising results when it comes to determining both conserved and novel copy number variations across tumor samples sequenced in a different time frame, as well as confidently finding variants depending on the characteristics of each caller and their combination. SB MultiCNV flexible design makes it suitable for a wide range of CNV analysis applications across the bioinformatics world.

Keywords:

bioinformatics, copy number variations, consensus calling, DNA sequencing

*Corresponding author, e-mail: jelisaveta.ilic@sbgenomics.com

Predicting suicide: serotonin presynapse dynamic modelling and machine learning approach

Lana Radenković[1*], Jelena Karanović[1], Maja Ivković[2,3], Aleksandar Damjanović[2,3], Maja Pantović-Stefanović[2] and Dušanka Savić-Pavićević[1]

[1] Center for Human Molecular Genetics, University of Belgrade- Faculty of Biology, Studentski trg 16, 11000 Belgrade, Serbia
[2] Clinic for Psychiatry, Clinical Centre of Serbia, Pasterova 2, 11000 Belgrade, Serbia
[3] University of Belgrade-Medical School, Doktora Subotića 8, 11000 Belgrade, Serbia

Abstract
To this day the best predictor of a suicide attempt is history of suicide attempts. Suicide prediction tools are scarce and mostly rely on subjective perception of the psychiatric symptoms and stressful life events questionnaires. In contrast, genetic variants are principally unchanging throughout an individual's life and could prove to be valuable predictors of suicide predisposition. In this study we explored the predictive potential of five selected genetic variants in serotonin system genes responsible for serotonin synthesis (TPH2), transport (SLC6A4) and degradation (MAOA).

Our study included 392 psychiatric patients with confirmed diagnosis of unipolar depression, bipolar disorder or schizophrenia, among which 172 attempted suicide. Using our genotypic data and kinetic coefficients from the literature, we developed a dynamic model of the serotonin presynapse to examine the immediate functional properties of genetic variants shown to influence gene expression levels. In the model it was assumed that the variants influence the amount of available synapse proteins, thus affecting the amount of available serotonin. We used three datasets with the outcome suicide attempt as input for supervised machine learning: model simulation data, patients' genotype dataset and their stressful life events questionnaire. The algorithms included were K-NN, Logistic regression, Naïve Bayes, Decision tree and SVM.

Statistical analyses of genotypic data and modelled serotonin concentrations did not reveal any significant differences between suicide attempters and non-attempters. However, machine learning prediction on model simulation data reached higher accuracy (71%) than only genotypic data (64%), while the highest prediction accuracy (76%) was reached using questionnaire data. Predictions made using combined questionnaire/genotype and questionnaire/model datasets were less accurate (55% and 62%, respectively). Despite the lack of statistical significance, the model generated more data per individual and was subsequently more informative in suicide attempt prediction than genotypic data solely. The model prediction accuracy was almost comparable to that of stressful life events data, which is currently the most commonly used suicide predictor. Our results imply that dynamic modelling of the serotonin system synapse that takes into account interpersonal variability expressed through genetic variation could contribute to explaining the individual predisposition for suicide, either as a risk or protective factor.

Keywords:
suicide, dynamic modelling, machine learning, serotonin system, genetic variants

*Corresponding author, e-mail: lana.radenkovic@bio.bg.ac.rs

Structural analysis of the interaction between the SARS-CoV-2 Spike protein and the human ACE2 receptor

M. Ghoula[1], S. Nacéri[1], S. Sitruk[1], D. Flatters[1], A-C Camproux[1], G. Moroy[1]

[1] *Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm,*
F-75013 Paris, France

Abstract

The year 2020 has been marked by the emergence of the highly pathogenic coronavirus SARS-CoV-2. SARS-CoV-2 has been rapidly and internationally spreading causing a serious global public health emergency, hence the importance of developing new drugs to inhibit the virus mechanism and to reduce global infection. The Spike protein, which is the key element for SARS-CoV-2 viral attachment, fusion and entry, is the main target for the development of antibodies, entry inhibitors and vaccines. The Receptor Binding Domain (RBD), which is located on the S1 subunit of the Spike protein, mediates viral entry through the Angiotensin Converting Enzyme 2 (ACE2) recognition. To develop anti-viral therapeutics for SARS-CoV-2, it is important to identify the amino acids stabilizing the SARS-CoV-2 RBD and ACE2 complex and to target specific regions of the complex in order to disrupt it.

In this aim, we focused our work on the interaction between the RBD and ACE2 receptor complex. Two crystallographic structures (pdb code : 6M0J and 6LZG) were used to understand the interaction mechanism of both proteins. Then, the complex and the isolated SARS-CoV-2 RBD protein's stability and flexibility were studied through Molecular Dynamics simulations with the GROMACS software. In total, we ran 20 simulations of 100ns each to cover a wide range of our different systems' conformational space. The free binding energy of the complex and the identification of contributing key hotspots were done using the Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) method. An extensive pocket search using the PockDrug software was also conducted to detect druggable pockets in the RBD protein. Altogether, our study helped us to identify interesting druggable pockets comprising crucial key residues for the RBD-ACE2 interaction and that can be targeted by efficient inhibitors that could potentially prevent the virus infection. Moreover, it has enlightened us about the new emerging mutations (K417N, N501T, E484K) impact and the understanding of their molecular mechanisms.

Keywords:
bioinformatics, structural biology, Covid-19

*Corresponding author, e-mail: mariem.ghoula@inserm.fr

De Novo Transcriptome Sequencing of *Ramonda serbica*: Identification of the Candidate Genes Involved in the Desiccation Tolerance

Marija Vidović [1*], Strahinja Stevanović [1], Sonja Veljović-Jovanović [2]

[1] *Institute of Molecular Genetics and Genetic Engineering, Laboratory for Plant Molecular Biology, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia*
[2]*Institute for Multidisciplinary Research, University of Belgrade, Kneza Višeslava 1, Belgrade, Serbia*

Abstract
Ramonda serbica Panc. is a resurrection plant that can survive a long period of severe dehydration-desiccation. Desiccation induces cellular membrane integrity loss, protein aggregation, and denaturation, as well as accelerated generation of reactive oxygen species. However, *R. serbica* can fully recover its metabolic functions already one day upon watering [1]. The aim of our study was to obtain more insight into the desiccation tolerance mechanisms by differential *de novo* transcriptomics of hydrated (HL) and desiccated leaves (DL). For *R. serbica* transcriptome construction, the total high-quality RNA from mixed samples of five biological replicates of HL and of DL separately, was extracted according to our previously optimised protocol [2]. Highly purified cDNA libraries were sequenced on an Illumina Hi-Seq platform. The ambiguous nucleotides, adapter sequences, and low-quality sequences were trimmed, and the quality of the reads was checked before and after the trimming. In total, 39608813 (with Q30=94%) and 37482969 (with Q30=94.1%) clean reads were obtained in HL and DL, respectively, and used to perform transcriptome assembly by Trinity software. After removing the redundancy, 189456 transcripts with 189003 unigenes were obtained (32.6% with the length between 500-1kbp).

Comparative analysis revealed that a large portion of *R. serbica* sequences (49.1%) exhibited high homology (according to obtained blast hits, e-value = 1e-5) with sequences found in the genome of another resurrection plant *Boea hygrometrica*. Furthermore, among the obtained unigenes (merged data for HL and DL), 64.6% and 42.3% were annotated by NCBI non-redundant protein and nucleotide sequences database (db), 23% by PFAM db, 22.5% by Clusters of Orthologous Groups of proteins db, 48.02% by Swiss-Prot db, 23 % KEGG db and 13.73 by Gene Ontology db. According to Blast2go analysis, the majority of annotated genes of *R. serbica* were associated with translation, ribosomal structure, posttranslational modifications, protein turnover, signalling pathways and cytoskeleton and encoded chaperonins and late embryogenesis abundant (LEA) proteins.

Aiming to provide a list of candidates involved in the desiccation tolerance in R. serbica we analysed differentially expressed genes in HL and DL. Genes associated with transmembrane transport, reproduction, cell proliferation, and protein folding were up-regulated in HL compared with DL. On the other hand, genes encoding proteins involved in cell wall architecture, LEA proteins and antioxidative defence were up-regulated in DL. Taken together, our results imply a key role of genes responsible for leaf morphological changes (wrapping and curling), and those encoding antioxidative enzymes (polyphenol oxidases and superoxide dismutases), as well as LEA proteins, known to be a hallmark of desiccation tolerance in resurrection plants.

*Corresponding author, e-mail: mvidovic@imgge.bg.ac.rs

Keywords:
antioxidative metabolism, differentially expressed gene analysis, drought, functional annotation, late embryogenesis abundant proteins, resurrection plants.

Acknowledgements:
This work was funded by the Science Fund of the Republic of Serbia-RS (PROMIS project LEAPSyn-SCI, grant no. 6039663), the Ministry of Education, Science and Technological Development, RS (Contract No. 451-03-9/2021-14/200042), Bilateral project (no. 451-03-01963/2017-09/09) and Centre for Green Technologies, University of Belgrade.

Benchmarking of Short and Structural Variant Calling Solutions available on SevenBridges Platform within GIAB Benchmarking Framework

Milica Aleksić[1*], Milan Kovačević[1], Jelena Ranđelović[1], Lea Lenhardt Acković[1] and Nevena Miletić[1]

[1] Seven Bridges, Schrafft's City Center, 529 Main St, Suite 6610
Charlestown, MA 02129, United States of America

Abstract
Recent advancements in sequencing technologies have enabled the development of new variant detection methods, which in order to transition from development to routine research and clinical practice require thorough benchmarking. Genome in a Bottle (GIAB) Consortium has developed a benchmarking framework for the identification of structural and small variants accompanied by the set of best practices to follow during the benchmarking procedure. This framework enables the detection of false positive and false negative calls within a defined confidence region taking into consideration differently represented variants calls, and reports results using a defined set of performance metrics. The aim of this project was to benchmark solutions for small and structural variant calling available on SevenBridges platform using a framework suggested by GIAB. The benchmarking was performed on data that leverages both short (Illumina) and long-read technologies (Pacific Biosciences HiFi and Oxford Nanopore PromethION). Sequencing data originated from HG002, Ashkenazi Jewish son from Personal Genome Project. The following solutions for processing short variants were benchmarked: (1) GATK Best Practice Whole Genome Germline SNPs and Indels Variant Calling Workflow (GATK version 4.1.0.0), (2) Longshot and (3) DeepVariant. For benchmark structural variant (SV) calling, following solutions were used: (1) ONT WGS Data Processing pipeline and (2) PacBio Whole Genome Alignment and SV Calling Pipeline. Both of these workflows include Sniffles, pbsv, cuteSV and SVIM as SV callers. The results were compared to the truth call sets defined by GIAB on GRCh38 and GRCh37.

Keywords:
long-read sequencing, GIAB, short variant calling, structural variant calling, benchmarking

*Corresponding author, e-mail: milica.aleksic@sbgenomics.com

Models for Prediction of Structural Alphabet Protein Blocks

Mirjana M. Maljkovic[1*], Nenad Mitic[1] and Alexandre G. de Brevern[2]

[1]*Faculty of Mathematics, University of Belgrade, Studentski trg 16,
11000 Belgrade, Serbia*
[2] *Université de Paris, INSERM UMR_S 1134, DSIMB, Université de la Réunion, INTS 6, rue Alexandre Cabanel 75015 Paris, France*

Abstract

One approach to approximation of 3D structure of the protein backbone is based on structural alphabets. Structural alphabets (SAs) are libraries that consist of patterns of local protein conformations. Patterns are determined based on fragments of consecutive amino acids in polypeptide chains. Protein Blocks (PBs) is one of the most known SA. PBs is composed of 16 patterns, labeled by letters ranging from PB a to PB p. Each protein block is defined with eight protein backbone dihedral angles of five consecutive amino acids. A sequence of PBs can be assigned to a protein chain with a known 3D structure. To each fragment of five consecutive amino acids in a sequence is assigned a PB with the lowest root mean square deviation on angular values. Besides the description of the 3D protein backbone, PBs are used to predict local structure from protein sequence. Several predictors have been developed for the assessment of the sequence of Protein Blocks for a given amino acid sequence. Their best-published accuracy is less than 70%. We developed models for PBs prediction using different machine learning algorithms in IBM SPSS Modeler. For the development of PBs prediction models, a dataset of 11,159 protein chains with a sequence identity of 25%, a resolution cutoff of 2.5Å, and a R-factor cutoff of 0.25 was used. Amino acid sequence, the output of the Spider3 predictor of the protein structure properties, the output of several predictors of the protein intrinsically disordered regions, and appearances of direct and inverse maximal repeats in protein sequence for each chain in the used dataset were combined and used as input for building PBs prediction models. The best accuracy of obtained PBs models is 80%, and it was achieved with the model developed using the C5.0 algorithm. In the analysis of the results of PBs models developed by applying different algorithms, it was noticed that properties of five amino acids that define a protein block are sufficient for the prediction of a PB, and the significance of the input attributes differs among the models developed by different algorithms.

Keywords:

Protein Blocks, prediction models, amino acid sequences, Spider3, repeats, IDR predictors

*Corresponding author, e-mail: mirjana@matf.bg.ac.rs

Polyploidy-induced atavistic regression to unicellularity and developmental bivalent gene activation in the context of carcinogenesis.

Ninel Miriam Vainshelbaum[1,4] , Olga Anatskaya[2], Alexander Vinogradov[2], Alessandro Giuliani[3], Jekaterina Erenpreisa[1*]

[1] Department of Oncology, Latvian Biomedical Research and Study Centre, Cancer Research Division, LV-1067 Riga, Latvia
[2] Department of Bioinformatics and Functional Genomics, Institute of Cytology, Russian Academy of Sciences, 194064 St. Petersburg, Russia
[3] Istituto Superiore di Sanità, 00161 Rome, Italy
[4] Faculty of Biology, University of Latvia, LV-1586 Riga, Latvia

Abstract
Recently, polyploidy has been established as an important driving force of cancer development, aggressiveness, and adaptation to treatment. Malignant tumors have also been shown to undergo atavistic regression – a phylostratigraphic and phenotypic shift from multicellularity to unicellularity. In order to investigate a possible link between these phenomena, the transcriptomes of polyploid and diploid human and mouse tissues (liver and heart) were compared using pairwise cross-species transcriptome analysis, principal component analysis and protein-protein interaction network analysis.

The results suggest that polyploidy causes the evolutionary age of the transcriptome to shift towards the more ancient unicellular and early metazoan phylostrata, upregulating unicellular metabolic and drug resistance-related pathways and downregulating circadian regulators. Furthermore, polyploidy was observed to upregulate bivalent genes, many of them involved in developmental processes and the c-MYC interactome. Investigation of the protein-protein interaction network of ploidy-activated bivalent c-MYC interactants demonstrated the involvement of gene hubs engaged in both embryonic development and carcinogenesis (including proto-oncogenes). Finally, tumor suppressor genes were shown to be downregulated.

These results reaffirm the importance of polyploidy in cancer development and evolution, as it appears to go hand-in-hand with atavistic regression, driving it epigenetically through the activation of developmental bivalent genes, and is overall able to promote an environment that benefits cancer development independently of mutations and chromosomal instability. The conclusions drawn from analyzing normal diploid and polyploid tissues warrant further (currently ongoing) validation in malignant tumor samples and may have potential diagnostic and/or prognostic importance.

Keywords:
transcriptomics, polyploidy, pairwise cross-species transcriptome analysis, cancer, bivalent genes, atavistic regression

*Corresponding author, e-mail: katrina@biomed.lu.lv

Database Construction and Network Analysis of Host-Pathogen Protein-Protein Interactions Involved in Microbial CVDs

Nirupma Singh[1], Sonika Bhatnagar[1]

[1]*Netaji Subhas Institute of Technology, Azad Hind Fauj Marg, Dwarka, New Delhi, India,*

Abstract

Microbial cardiovascular disease (CVD) is a less explored class of CVDs. Many host pathogen protein-protein interactions (HP-PPIs) are found to involved in microbial CVDs which led to the construction of an online database named MorCVD. Network biology approach was chosen to get a summarized view of the HP-PPIs. A tripartite network of pathogens involving pathogenic proteins and host proteins was constructed. Topological and functional analysis of the network was carried out to establish a correlation between both. Topological analysis was done to look for the centrally located proteins and pathogens in the network. Biological characterization was performed to look for the biological nature of the central proteins. Ontology annotation and pathway analysis allowed further insight into underlying pathophysiology and mechanism employed by proteins to lead towards CVD from infection.

From the analysis, it has been observed that viral pathogens interact more with the human proteins as compared to bacterial pathogens. Essential host factors and central host proteins of the network were over-represented amongst the subset of host proteins interacting with central viral proteins. The central and the virulent proteins of the bacteria interact with a much smaller number of host proteins in comparison with virus. The main pathways leading from organism-specific infections to microbial CVDs are NF-κB signaling pathway, Toll-like receptor signaling pathway, TNF signaling pathway and T cell receptor signaling pathway. Some domains of pathogen proteins were found to mimic the host protein domains to hijack the host machinery. The gene ontology, pathway analysis and significant enrichment of immune-related proteins amongst the central nodes point towards the significance of immune response in CVD effects of microbial infections.

Keywords:

Tripartite network, Topological analysis, Ontology analysis, Biological characterization, Central proteins

*Corresponding author, e-mail: sbhatnagar@nsut.ac.in

A Subtractive Proteomics Approach to Identify Putative Drug Targets Against Parasitic Species

Parakh Sehgal [1], Shradha Mahatre [2]

[1] Biotechnology Department, G D Rungta College of Science and Technology, Bhilai, Durg, Chhattisgarh, 490024, India.
[2] Biotechnology Department, Patkar-Varde college, Goregoan (W), Mumbai-400062, India.

Abstract

Parasitic diseases affecting humans continue to be the leading cause of morbidity as well as mortality, particularly in the tropical and subtropical countries. It's a need of hour to develop new strategies for prediction of novel drug targets with increase of drug/multi-drug resistant forms of many diseases. Accessible proteome sequence data of pathogens have provided voluminous information that can be useful in identification of ideal drug target. A known method subtractive proteomics approach (SPA) used for subtraction of sequence dataset between the host-pathogen proteome and provides data pertaining to a set of proteins that are likely to be essential to the pathogen but absent in the host. This work is based on identification of new drug targets of parasitic species namely *Leishmania major, Plasmodium falciparum, Trypanosoma brucei and Trypanosoma cruzi* using SPA method as an efficient one.

Material and methods

Transporter Database 2.0 was used to predict common proteins of mentioned parasitic species. Their FASTA file were retrieved from Uniprot database and subjected to BlastP tool to discovery proteins which are non-homologous to Humans. Further Database of Essential Genes (DEG) was used to investigate non-homologous sequences to shortlist essential proteins. Use of CELLO tool detected sub-cellular location of essential proteins and conserved domains were identified using PROSITE Database.

Results

Transporter 2.0 predicted 22 common essential proteins (17 plasma membrane, 3 mitochondrial and 1 inner membrane, 1 nuclear membrane) respectively. The analysis predicted novel drug target common in mentioned species as mitochondrial Carrier (MC) family, Major Facilitator Superfamily (MFS), Sulfate Permease (SulP) Family, P-type ATPase (P-ATPase) Superfamily and Major Intrinsic Protein (MIP) Family proteins.

Key words

SPA-Subtractive proteomics approach, DEG- database of Essential Genes

*Corresponding author, e-mail: katrina@biomed.lu.lv

The possible role of hydrogen peroxide molecules in ion beam therapy of cancer cells

Sergey N. Volkov

Bogolyubov Institute for Theoretical Physics,
National Academy of Sciences of Ukraine,
03143 Kiev, Ukraine

Abstract

To date one of the most promising treatments of cancer diseases is radiation therapy. In this approach the tumor is irradiated with a beam of particles (protons, electrons, neutrons, ions, etc.), which leads to its destruction. Irradiation with protons and heavy ions is most effective because these particles, passing through the medium, lose most of their energy in a certain small area inside the body. Ion therapy is now considered as effective and safe method of cancer treatment. But the processes leading to the deactivation of cancer cells during ion therapy are still poorly understood. In particular, the role of hydrogen peroxide molecules, which are formed in cells during irradiation, is practically unknown.

In the presented work the possible mechanisms of action of hydrogen peroxide molecules on DNA functioning in biological cell are studied. Using quantum-mechanical approach the competitive interaction of hydrogen peroxides and water molecules with DNA recognition sites is analyzed and the advantage of hydrogen peroxide molecules in the formation of complexes with DNA nucleic bases is shown. According to our calculations, the binding energy between atomic groups of DNA and hydrogen peroxides is always greater than the same energy of interaction with water molecules. The estimation of lifetimes of the complexes of peroxide or water molecules with DNA atomic groups shows the possibility of blocking of DNA genetic activity by peroxide molecules. So, the lifetime of a complex of guanine base of DNA with peroxide molecule is more than 50 times larger than with water, and for adenine base - more than 30 times. This advantage is explained by the formation of stronger hydrogen bonds between the peroxide molecules and nucleic bases. Thus, as a result of cell irradiation, long-lived molecular complexes of DNA nucleic bases with hydrogen peroxide molecules are formed, which should lead to blocking the processes of DNA genetic activity in the cancer cell as a whole.

Keywords:

ion beam therapy, DNA, hydrogen peroxide

*Corresponding author, e-mail: snvolkov@bitp.kiev.ua

A bioinformatics approach to inferring environmental drivers of SARS-CoV-2 transmissibility

Sofija Marković[1*], Ognjen Milicevic[2], Andjela Rodic[1], Dusan Zigic[3], Marko Tumbas[1], Igor Salom[3], Magdalena Djordjevic[3], Marko Djordjevic[1]

[1]Quantitative Biology Group, Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Serbia
[3]Department for Medical Statistics and Informatics, School of Medicine, University of Belgrade, Serbia
[2]Institute of Physics Belgrade, National Institute of the Republic of Serbia, University of Belgrade, Serbia

Abstract

We here present a novel approach we developed to infer the main factors of COVID-19 transmissibility, which overcomes a number of existing difficulties: dependence of case counts on testing-policy and intervention measures, high dimensionality, and multicollinearity of predictors. As a measure of the disease transmissibility, independent of intervention measures and testing policies, we inferred SARS-CoV-2 basic reproduction number (R_0) for 118 world countries and 46 USA states. For these regions, we assembled a broad spectrum of sociodemographic, weather and health-related covariates. We performed principal component analysis (PCA) on conceptually similar subsets of these variables. This feature engineering reduces the dimensionality and collinearity of the predictors, while achieving straightforward PC interpretability. We then used linear regressions, which provide both regularization and variable selection (Lasso and Elastic Net). Also, to estimate variable importance and quantitate their effect on R_0, we used ensembles of decision trees (Gradient Boost and Random Forest). On a global level, we obtained that country's prosperity (GDP per capita or Human Development Index) is the main predictor of COVID-19 spread, likely as a good proxy for the social contact frequency. Additionally, there are important contributions from unhealthy living conditions and lifestyle (manifested mainly through obesity, physical inactivity, smoking and air pollution), which promote transmissibility. For the USA, variations of sociodemographic factors are much smaller than on a global scale, so we obtained ambient air pollution, in particular $PM_{2.5}$ concentration, as the main driver of COVID-19 transmissibility. These results address a challenging yet crucial task to help understand, predict, and potentially prevent future COVID-19 outbursts.

Keywords:
COVID-19 transmissibility, basic reproduction number, air pollution, COVID-19 demographic dependence, principal component analysis, machine learning

*Corresponding author, e-mail: sofija.markovic@bio.bg.ac.rs

Using human genetics to understand the aetiology of reproductive ageing and its links to later life diseases

Stasa Stankovic

University of Cambridge, MRC Epidemiology, Addenbrooke's Hospital,
Cambridge, United Kingdom

Abstract

Reproductive longevity is critical for fertility and impacts healthy ageing in women, yet insights into the underlying biological mechanisms and treatments to preserve it are limited. The current markers used in clinical practice are not precise predictors of the menopause timing, thus not allowing informed reproductive choices for women and early treatments. By combining state-of-the-art genomic technologies with human population-based studies, including UK Biobank, Reprogen consortium and deCODE Genetics, we identified 290 common genetic variants through genome-wide association study (GWAS) that govern ovarian ageing assessed using normal variation in age at natural menopause (ANM) in - 200,000 women of European ancestry. These common alleles influence clinical extremes of ANM; women in the top 1% of genetic susceptibility have an equivalent risk of premature ovarian insufficiency to those carrying monogenic FMR1 premutations. As fertility declines 10 years prior to menopause and 1 in 100 women have menopause before the age of 40, prediction of women with the most risk of infertility due to premature ovarian ageing is essential. Our current genetic risk scores using identified loci outperform other non-genetic causes but are not yet at the level where they can be deployed clinically (AUC-ROC ~0.65).

Identified loci implicate a broad range of DNA damage response (DDR) processes and include loss-of-function variants in key DDR genes identified through analysis of UK Biobank whole exome sequencing data. We demonstrate that experimental manipulation of DDR pathways highlighted by human genetics increases fertility and extends reproductive life in mice, acting across the life-course to shape the ovarian reserve and its rate of depletion. Causal inference analyses using identified genetic variants indicates that extending reproductive life in women improves bone health and reduces risk of type 2 diabetes, but increases risks of hormone-sensitive cancers. Our current work involves use of large-scale proteomic and metabolomics data, as well as exome level data to identify low-frequency and rare variants not well captured by current array-based data. Our findings provide insight into the mechanisms governing ovarian ageing and how they might be targeted by therapeutic approaches to extend fertility and prevent disease.

Keywords:

GWAS, causal inference, population studies, whole exome sequencing, genetic risk score, ovarian ageing, DNA damage response

*Corresponding author, e-mail: stasa.stankovic@mrc-epid.cam.ac.uk

Hierarchical Protein Function Prediction with Tail-GNNs

Stefan Spalević[1*,3], Petar Veličković[2], Jovana Kovačević[1], and Mladen Nikolić[1]

[1] *Faculty of Mathematics, University of Belgrade, Studentski trg 16,*
11000 Belgrade, Serbia
[2] *DeepMind, 6 Pancras Square, London N1C 4AG, UK*
[3]*School of Electrical Engineering, University of Belgrade, Bulevar Kralja Aleksandra 73,*
11000 Belgrade, Serbia

Abstract

Knowing the function of a protein informs us on its biological role in the organism. With large numbers of genomes being sequenced every year, there is a rapidly growing number of newly discovered proteins. Protein function is most reliably determined in *wet lab* experiments, but current experimental methods are too slow for such quick income of novel proteins. Therefore, the development of tools for automated prediction of protein functions is necessary. Fast and accurate prediction of protein function is especially important in the context of human diseases since many of them are associated with specific protein functions.

The space of all known protein functions is defined by a directed acyclic graph known as the Gene Ontology (GO), where each node represents one function and each edge encodes a hierarchical relationship between two functions, such as is-a or part-of. For every protein, its functions constitute a subgraph of GO, consistent in the sense that it is closed with respect to the predecessor relationship. GO contains thousands of nodes, with function subgraphs usually having dozens of nodes for each protein. Hence, the output of the protein function prediction problem is a subgraph of a hierarchically-structured graph.

Graph neural networks (GNNs), with their built-in inductive bias for relational data, are therefore naturally suited for this task. However, in contrast with most GNN applications, the graph is not related to the input, but to the *label* space. Accordingly, we propose *Tail-GNNs*, neural networks which naturally compose with the output space of any neural network for multi-task prediction, to provide relationally-reinforced labels. For protein function prediction, we combine a Tail-GNN with a dilated convolutional network which learns representations of the protein sequence, making significant improvement in F_1 score and demonstrating the ability of Tail-GNNs to learn useful representations of labels and exploit them in real-world problem solving. Due to limitations of computational resources we used reduced ontology with 123 nodes which represent protein functions and 145 edges for their relations. On the reduced ontology of Molecular Functions in GO, the proposed Tail-GNN model with sum aggregation and spectral features achieved F_1 score 0.6 which represents an improvement in comparison to baseline network which does not use graph neural networks and achieved F_1 score 0.584.

Keywords:
protein function prediction, graph neural networks

*Corresponding author, e-mail: stefan.spalevic@etf.bg.ac.rs

Imidazolines: *In silico* off-target fishing in the class A of G protein-coupled receptors

Teodora Djikic[1*], Jelica Vucicevic[1], Jonne Laurila[2], Marco Radi[3], Nevena Veljkovic[4], Henri Xhaard[5] and Katarina Nikolic[1*]

[1] *Department of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Belgrade, Vojvode Stepe 450, 11000 Belgrade, Serbia.*
[2] *Research Center for Integrative Physiology and Pharmacology, Institute of Biomedicine, University of Turku, Turku, Finland*
[3] *Dipartimento di Scienze degli Alimenti e del Farmaco, Università degli Studi di Parma, Viale delle Scienze, 27/A, 43124 Parma, Italy*
[4] *Division of Pharmaceutical Chemistry, Drug Research Program, Division of Pharmaceutical Chemistry and Technology, Faculty of Pharmacy, University of Helsinki, P.O. Box 56, FI-00014 University of Helsinki, Helsinki, Finland*
[5] *Laboratory for bioinformatics and computational chemistry, Institute of Nuclear Sciences Vinca, University of Belgrade, Mihaila Petrovica Alasa 14, 11001 Belgrade, Serbia*

Abstract

Centrally acting hypotensive imidazoline derivatives are agonists of α_2-adrenoceptors and non-adrenergic I1-imidazoline receptors. Based on the finding that a central antihypertensive agent with high affinity for I1-type imidazoline receptors – rilmenidine, shows cytotoxic effects on cultured cancer cell lines, it has been suggested that imidazoline receptors agonists might have a therapeutic potential in cancer therapy. Nevertheless, rilmenidine itself does not represent a suitable candidate because of possible side effect caused by activation of α_2-adrenergic receptors. In our previous work, several novel rilmenidine-derived compounds with anticancer potential and without an agonistic activity on α_2-adrenoceptor were identified. Taking into consideration that human α_2-adrenergic receptors belong to the rhodopsin-like class A of G protein-coupled receptors (GPCRs), the biggest group of drug targets, that share structure similarity, it is reasonable to assume that these ligands might have the affinity on some other receptors from the same class.

To investigate potential additional targets for novel imidazoline I1 agonists a reverse docking protocol on 107 GPCRs, using 63 imidazoline ligands and their 670 decoys was prepared. Unlike typical molecular docking protocol, where series of small molecules are docked in one macromolecular target, in case of reverse docking a small-molecule is docked in the set of potential target proteins. Due to the availability of crystal structures all of the included GPCRs, were in inactive state, and therefore suitable for identification of the receptors antagonized by imidazoline ligands. Based on ROC curves and Enrichment Factors, 20 potential off-target GPCRs were selected. To better assess the affinity of imidazoline derivatives for chosen receptors, an additional docking study was performed, and docking scores of imidazolines were compared with docking scores of known antagonists. Finally, to verify *in silico* results, three ligands with high scores and tree ligands with low scores were tested for antagonistic activity on α_2-adrenergic receptors.

The protocol described here could be applied on all the small molecules, for the detection of potential interactions with GPCRs of class A, in the early stage of drug design process. Additionally, this protocol is easily expandable, by adding novel receptors/subfamily of receptors as soon as the crystal structures and/or 3D models become available.

Keywords:

off-target, cross-docking, GPCRs, imidazolines, adrenoceptors

*Corresponding author, e-mail: teodora.djikic@pharmacy.bg.ac.rs, katarina.nikolic@pharmacy.bg.ac.rs

Differential analysis of co-expression networks in the dog mammary gland carcinomas

Teodora Lukić[1*], Bogdan Jovanović[1], and Vladimir M. Jovanović[2,3]

[1] *Center for Human Molecular Genetics, Faculty of Biology, University of Belgrade, Studentski trg 3, 11000 Belgrade, Serbia*
[2] *Bioinformatics Solution Center, Freie Universität Berlin, Takustraße 9, 14195 Berlin, Germany*
[2] *Human Biology and Primate Evolution Group, Freie Universität Berlin, Königin-Luise-Straße 1-3, 14195 Berlin, Germany*

Abstract
With more than 2 million cases per year, breast cancer is the most frequent type of malignancy and one of the leading causes of death in women. When detected early, it is curable at relatively high success, although the process of treatment negatively impacts the life quality of a patient and family. As many different molecular subtypes are present, it is very important to further understand the biology of breast cancer in order to develop better diagnostics and treatments.

Due to numerous known similarities between human and canine mammary tumors, dogs with mammary carcinomas could be model organisms for the study of human breast cancer. A better understanding of gene networks involved in the canine mammary tumors could shed some light on the basis of the grave human disease. To this aim, the meta-analysis of several published transcriptomic (RNA-Seq) experiments of normal and tumor dog mammary tissues was conducted. After the differential expression analysis using the DESeq2 pipeline was performed, the obtained differentially expressed genes were identified as nodes of further interest in the co-expression network analysis. The overlapping weights of these nodes were computed in the signed weighted topological overlap analysis (wTO), and the consensus co-expression networks were created for all subsamples. The systematic comparison of retrieved consensus networks detected common, specific and different nodes (differentially expressed genes) and links (co-expression weights) for healthy and tumor tissue. The results of gene ontology and pathways enrichment analysis within the node groups were compared to the studies of genes connected to human breast cancer.

Keywords:
breast cancer, differential gene expression, wTO, CoDiNa, co-expression

*Corresponding author, e-mail: m1044_2020@stud.bio.bg.ac.rs

Altered splicing profile of Ptbp1 drives global splicing tune-up upon neuroinflammation response in rat model

Vladimir Babenko[1*], Galina Shishkina[1], Dmitri Lanshakov[1], Tatiana Kalinina[1], Elena Sukhareva[1] and Nikolay Dygalo[1]

[1] *Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia*

Abstract

Neuroinflammation is considered as a key pathological event that causes the development of cognitive impairments and psychiatric disturbances associated with many neurodegenerative diseases, including post-ischemic consequences. In order to address them, lipopolysaccharide (LPS) administration is used as an effective approach to model the inflammatory state In our study, LPS was infused into the striatum, the brain region rapidly affected by ischemia followed by hippocampus (HIPP), playing role in regulation of memory and emotion. Distant damage of this structure may be implicated in long-term negative effects.

We used HIPP transcriptome (RNA-Seq) analysis of processes emerged upon administering LPS while using Saline (SAL) injected animals, three samples per group as a control. We collected the samples after 24 hours. We explored the Alternative splicing (AS) in HIPP of SAL-LPS rat samples to elucidate peculiarities of AS events. We observed around 10 000 AS cases by exon skipping (ES). From them 68 differential alternative splicing exon skipping events (FDR<0.1) in the samples of HIPP vs SAL were observed.

In particular, we observed a distinct significant *Ptbp1* master regulator splicing pattern alteration by excluding exon 8 (Gueroussov et al., 2015; PMID: 26293963), followed by modified Ptbp1 itself essentially altering the splicing pattern of approximately 1,5 thousands exons in various genes (Gueroussov et al., 2015). We speculate that neuroinflammation could be the prime cause of this particular observation, which may also lead to excessive glia genesis. Notably, the similar *Ptbp1* splicing alteration effect was observed in type 1 mice diabetic hearts (Belanger et al., 2019; PMID: 30594394).

The report will outline the differential splicing events in target genes set, along with the correspondent networks, which may prove the point of profound changes upon neuroinflammation resulting in long lasting alternative splicing alteration in adult brain. We underline that the intense splicing plasticity is manifested by specific postsynaptic genes as well as immune response genes including interferon system and von Willebrand factor related genes. We also assessed AS-mediated proteome diversity expansion by considering exon insertion/skipping alterations tune-up while keeping coding potential valid.

Keywords:

transcriptome sequencing, Alternative splicing, animal model, ischemia, neuroinflammation

Acknowledgment:

The work was funded with Russian Science Foundation N 20-64-47013.

*Corresponding author, e-mail: bob@bionet.nsc.ru

E-boxes as ZEB2 binding sites

Vladimir M. Jovanovic[1,2*], Jeong-Eun Lee[1], Amanda Jager Fonseca[1], Sebastian Streblow[1], Stefano Berto[3], and Katja Nowick[1]

[1] *Human Biology and Primate Evolution Group, Freie Universität Berlin, Königin-Luise-Straße 1-3, 14195 Berlin, Germany*
[2] *Bioinformatics Solution Center, Freie Universität Berlin, Takustraße 9, 14195 Berlin, Germany*
[3] *Medical University of South Carolina, Charleston, SC, USA*

Abstract

One of KRAB-ZNF proteins, the zinc finger E-box binding homeobox 2 (ZEB2), has previously been revealed as one of the gene regulatory factors with the largest number of human-specific links in the prefrontal cortex, in contrast to chimpanzees. The change of its function due to mutations leads, among other systemic problems, to cognitive disabilities seen in a rare Mowat-Wilson syndrome. This makes ZEB2 an ideal candidate for human evolution studies, where the uncovering of its dissimilar functions in different primate species could add to our knowledge of the regulation of human-specific, especially brain-related traits.

ZEB2's thoroughly investigated binding to the enhancer boxes (E-boxes) in a promoter region of E-cadherin was seen as very important in defining its gene regulatory function, consequently even becoming eponymous. In order to analyse ZEB2 regulatory function better, ChIP-Seq study was performed in triplicate on B-lymphoblastoid cell lines from human, chimpanzee and orangutan donors. The binding sites were determined by MACS2 peak calling procedure, and after the quality assessment kept if present in at least two samples of a species. The presence of E-boxes within recovered binding sites was then analysed, with the accent put on the frequency of ZEB2-binding tandem E-box sequences known from the E-cadherin promoter (CACCTG/CAGGTG). Surprisingly, other E-box sequences were more common in the studied samples, with a sharp difference between the human and orangutan ones. The architecture of binding sites and the evolutionary significance of the discovered differences are discussed.

Keywords:

DNA binding site, ChIP-Seq, transcription factor, human evolution, ZEB2

*Corresponding author, e-mail: vladimir.jovanovic@fu-berlin.de

Innovative bioinformatic approach to kidney transplant wait-list management in the Republic of Serbia

Vladimir Perovic[1*], Nikola Zogovic[2] and Nevena Zogovic[3]

[1] Institute of Microbiology and Immunology, Medical Faculty University of Belgrade, Dr Subotica 1, Belgrade, Serbia
[2] Institute Mihajlo Pupin, Volgina 15, Belgrade, Serbia
[3] Department of Neurophysiology, Institute for Biological Research "Sinisa Stankovic" – National Institute of Republic of Serbia, University of Belgrade, Bulevar despota Stefana 142, 11000, Belgrade, Serbia

Abstract
Renal failure represents a growing clinical problem around the world. Although dialysis is a short-term solution, kidney transplantation confers better survival and quality-of-life outcomes for most patients with end-stage kidney disease. A major limitation to renal transplantation is the supply of donor kidneys. Determining eligibility for a kidney transplantation is one of the most difficult decisions facing clinicians. Clinical practice guidelines have been implemented in many countries for the evaluation and acceptance of patients for the kidney transplantation waiting list in order to provide explicit recommendations to guide clinical decision making. The most important determinants of the outcome of renal transplantation are the degree of HLA matching, the cold ischemia time (total time between removal of the kidney from the donor and its transplantation into the recipient), blood group matching, number of prior grafts, presence of donor-specific antibodies, age of donor and recipient, time on dialysis prior to transplantation, diabetes in the recipient, race, living or cadaver donor, and transplant center. Kidney allocation algorithms vary both within and between countries. Most methods of donor organ allocation involve the use of simple algorithms designed to take into account major factors thought to influence graft outcome.

The aim of this study is to improve the existing decision support system for Kidney Exchange Program (KEP) in the Serbian healthcare system by applying complex multicriteria optimization (CMCO) methods. Also, the goal is to determine the framework for harmonization of KEP in Serbia with the corresponding programs in European countries. In this study, we present the objectives and constraints in the Serbian KEP, determined by the medical aspects of kidney transplantation process and the Serbian law on human organ transplantation. We will then compare them with the corresponding KEP objectives and constraints in European countries. Based on the comparison and analysis of the applied CMCO algorithms in European countries with developed KEP, we intend to determine the guidelines for the CMCO algorithm in the Serbian KEP.

Keywords:
bioinformatics, kidney transplantation, kidney allocation, multicriteria optimization

*Corresponding author, e-mail: vladimir.perovic@med.bg.ac.rs

POSTER PRESENTATIONS:

TH Simulation Pipeline - Workflow for *in silico* Simulation of Tumor Samples with Known Tumor Purities

Milica Aleksić[1*], Ajša Nuković[1*], Luka Topalović[1] and Vojislav Varjačić[1]

[1] Seven Bridges, Schrafft's City Center, 529 Main St, Suite 6610
Charlestown, MA 02129
United States of America

Abstract
Sensitivity of methods that assess the variation in a tumor depend greatly on the fraction of normal cells in a tumor sample. Thus, estimating tumor purity of a sample plays an essential role in cancer research and diagnostics. There are multiple community- developed tools that tackle this problem, however development and verification of new methods relies on proper benchmarking methodology. Obtaining truth purity values for tumor samples is a necessity in that process, and usually relies on histopathological estimations which are often imprecise, especially going towards lower purity values. We present a workflow designed to simulate tumor samples with different purity values, starting from whole exome sequencing (WES) tumor and normal BAM files. The workflow assumes that tumor and normal samples are high purity, and simulates nine new tumor samples with purities ranging from 1 to 99% depending on user defined settings, while keeping all samples with uniform preset coverage. The workflow is composed of a BAM subsampling section (samtools view), subsampled BAM files merging section (bamtools merge) and postprocessing and quality check (QC) section (samtools sort, samtools index and Picard CollectHsMetrics). The workflow was validated by comparing purity estimates by different tumor purity estimation tools on in vitro and in silico simulated data, and our results show that estimates of multiple tools differ on average by 6% in their estimations, thus confirming the reliability of the pipeline. The workflow was further used on samples originating from different cell lines, cancer patient samples and The Cancer Genomics Atlas (TCGA) data in order to benchmark different tools for tumor purity estimation.

Keywords:
bioinformatics, simulation, tumor purity

*Corresponding author, e-mail: milica.aleksic@sbgenomics.com

Codon Usage-based SARS-CoV-2 protein classification

Aleksandar Veljković[1], Biljana Stojanović[2], Saša Malkov[1], Miloš Beljanski[3], Gordana Pavlović-Lažetić[1], and Nenad Mitić[1*]

[1]Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[2]Mathematical Institute of the Serbian Academy of Sciences and Arts, Kneza Mihaila 36, 11000 Belgrade, Serbia
[3]Institute for General and Physical Chemistry, Studentski trg 12, 11000 Belgrade, Serbia

Abstract
Severe acute respiratory syndrome corona virus 2 (SARS-CoV-2) appeared in late 2019 and spread across the world causing pandemic in humans. Since viruses differ in their specificity toward host organisms, analysis of the viral genome organization contributes to better understanding of their evolution and adaptation in the host. Polymorphism in genomic composition is reflected in its codon and amino acid usage patterns, as well as in translation rate (where rare codons are assumed to be translated more slowly than common codons). The same holds for specific coding sequences and the corresponding types of proteins. The goal of the current research is to build a model for classification of proteins (or parts thereof) based on codon usage patterns.

As a dataset we used the NCBI dataset of all the SARS-CoV-2 isolates and their coding sequences, preprocessed as to eliminate those with missing values, ambiguous letters and full duplicates, ending up with around 66000 isolates and around 770000 coding sequences (ORFs). We performed cluster analysis of all the coding sequences from the dataset based on codon usage (CU) as a means of identification of number and "profile" of protein classes. The approach is sound since the external as well as internal measures of clustering quality are high.

Results of clustering using TwoStep algorithm (using IBM SPSS Modeler tool) include 12 clusters containing almost perfectly separated types of proteins. This may be used as an argument that specific types of proteins have their specific codon usage patterns which may be then used for protein classification model.

Except for classification model based on protein clustering, we experiment with clustering virus isolates by following dynamics of CU patterns as a function of time during pandemic.

Keywords:
SARS-CoV-2, codon usage, CU, protein classification, protein clustering

*Corresponding author, e-mail: nenad@matf.bg.ac.rs

Prediction of protein function using binary classification methods

Anja Bukurov[1], Jovana Kovačević[1]

[1] *Faculty of Mathematics, University of Belgrade, Studentski trg 16,*
11000 Belgrade, Serbia

Abstract

Protein function is determined experimentally, which is an expensive and slow process. With the increasing amount of newly discovered proteins due to a large number of sequenced genomes, automated function prediction has become necessary.

The protein function may be represented by a directed acyclic graph, defined through The Gene Ontology project. Each node in this graph represents a function which is more general than its children's protein functions. Therefore, prediction of a protein function comes down to finding a subgraph of the Gene Ontology with the following consistency requirement: if node v is in the predicted subgraph, then the subgraph must contain all ancestors of v up to the root. Due to the limited number of proteins with specific functions (insufficient for training and testing) and to the fact that some nodes were marked as obsolete, the ontology has been reduced to 399 nodes.

We developed three machine learning models based on the following binary classification methods: SVM (with Radial basis function kernel), Random Forest (with 100, 400, 700 and 1000 trees) and Logistic Regression (with different values for parameter C). In every model, each of 399 protein functions was predicted individually. In order to obtain the complete function of a protein, for each model, the 399 binary predictors' results were combined into one by joining their answers manually, providing three different predictions.

We used protein sequences from UniProt, a total of 20960 proteins from different organisms with at least one of the 399 functions. Sequences were represented using a frequency array of trigrams of the 20 standard amino acids. For evaluation we used the test set with proteins of different species including *Human*, *Mouse*, *Rat*, *E. Coli* and *Arath*. Each of the predictors was tested on the test set containing proteins of different species as well as on its subsets containing proteins of single species. On the multi-species test set, all models had similar F_1-score. As for single species subsets, all models had slightly better F_1-score on *Human*, *Mouse* and *Rat* proteins than on *E. Coli* and *Arath* (except for the Random Forest for *E. Coli* which was better).

Keywords:

protein function, primary structure, binary classification methods, SVM, Random Forest, Logistic Regression

*Corresponding author, e-mail: anja_bukurov@matf.bg.ac.rs

Clustering of Transcription Factor Binding Sites in Model Plant Genomes

Yuriy L. Orlov[1,2*], Arthur I. Dergilev[1], Oxana B. Dobrovolskaya[1,3]

[1] *Novosibirsk State University, Pirogova, 1, 630090 Novosibirsk, Russia*
[2] *The Digital Health Institute, I.M.Sechenov First Moscow State Medical University (Sechenov University), Trubetskaya 8-2, 119991 Moscow, Russia*
[3] *Peoples' Friendship University of Russia (RUDN University), Miklukho-Maklaya str.6, 117198 Moscow, Russia*

Abstract

Recent advances in high-throughput sequencing technologies allow analysis the transcription factors binding in genome scale. The clusters of transcription factor binding sites determine regulatory gene regions. It defines evolutionary patterns of gene regulation changes in the species. The growth of the data volume on the experimentally found binding sites (ChIP-seq and related technologies) raises qualitatively new problems for the analysis of gene expression in model genomes. However, such data were not investigated in plant genomes in detail comparing to mammalian genomes. Plant genomes remain an insufficiently studied object, although they have complex molecular regulatory mechanisms of gene expression and response to the environmental stresses.

It is important to develop new software tools for the analysis of the transcription factor binding sites location, their clustering in a model genome, visualization, and statistical estimates for such clusters. The statistics of non-random clusters of the binding sites for 3 and more different factors identified by ChIP-seq was shown previously. Such clusters of sites could be used for gene promoter and enhancer prediction.

This work presents a new application for the analysis of transcription factor binding sites in several evolutionarily distant model plant organisms. We developed computer scripts to analyze ChIP-seq data, automatically define clusters and visualization it in the heatmaps format. We used ChIP-seq profile peaks to study the transcription factor binding in three plants, including *Arabidopsis thaliana, Physcomitrella patens,* and *Chlamydomonas reinhardtii.*

We discussed statistical estimates of the binding sites clustering. The non-random clusters of binding sites in the plant genomes were constructed. The transcription binding clusters in Arabidopsis were considered in more details including gene ontology and functional annotation.

Keywords:
bioinformatics, transcription factors, plants, genomes, sequencing

*Corresponding author, e-mail: orlov@d-health.institute

De novo transcriptome assembly and characterization of the liver transcriptome of Binni (*Mesopotamichthys sharpeyi*) using RNA-Seq data

Ayeh Sadat Sadr[1*], Mohammad Taghi Beigi Nasiri[2]

[1] *South of Iran Aquaculture Research Institute, Iranian Fisheries Science Research Institute (IFSRI), Agricultural Research Education and Extension Organization (AREEO), Ahvaz, Iran*
[2] *Department of Animal Science, Faculty of Animal & Food Science, Khuzestan Agricultural Sciences and Natural Resources University, Mollasani, Iran*

Abstract
Mesopotamichthys sharpeyi Fish is a cyprinid species, known as Binii. This species is native freshwater fishes found in Iran, Iraq, Turkey and Syria. Binii has been recently considered as a proper candidate for aquaculture in Iran due to its high resistance to a wide range of the environmental conditions, having proper features for rearing and marketability. Genomic and transcriptomic information for this organism is extremely deficient. The aim of this study was to characterize *Mesopotamichthys sharpeyi* hepatic transcriptome using Illumina paired-end sequencing the de novo transcriptome assembly generated 131022 unigenes, with an average length of 594 bp, N50 of 874 bp and 70% of complete and single-copy core vertebrate genes orthologues. The analysis of species distribution revealed that *Mesopotamichthys sharpeyi* contigs had the highest number of hits to Sinocyclocheilus. A total of 36069 microsatellites were detected in this study by MISA. This study presents the first transcriptome resources for *Mesopotamichthys sharpeyi* and provides basic information for the more analysis on transcriptome, such as the identification of RNA markers.

Keywords:
Mesopotamichthys sharpeyi, Denovo, RNAseq

*Corresponding author, e-mail: a.sadr@areeo.ac.ir

Statistical Data Analysis on the Human and Ecological Toxicity of Polycyclic Aromatic Hydrocarbons

Ryung Kyung Lee[1,2] and Baeckkyoung Sung[1,3*]

[1] *KIST Europe Forschungsgesellschaft mbH, Campus E7 1, 66123 Saarbrücken, Germany*
[2] *Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, 02841 Seoul, Republic of Korea*
[3] *Division of Energy & Environment Technology, University of Science & Technology (UST), 217 Gajeong-ro, Yuseong-gu, 34113 Daejeon, Republic of Korea*

Abstract
The environmental impact of anthropogenic compounds has been extensively studied to reveal their influence on the public health and ecosystems. There are three main approaches for performing the chemical toxicity evaluations: *in vivo*, *in vitro*, and *in silico*. Among them, the *in silico* toxicity testing has recently emerged as the most preferred method because of its high cost-efficiency and robustness for high-throughput screening. Driven by strong demands from industries and governments, the development of open software tools for the *in silico* toxicity screening and prediction is now an active topic in the field of computational toxicology. The QSAR Toolbox, developed and distributed by the Organization for Economic Cooperation and Development (OECD), is the representative and easy-to-use software package that enables the automated data collection and processing for rapid prediction of chemical toxicity, including the framework of quantitative structure-activity relationship (QSAR).

The polycyclic aromatic hydrocarbons (PAHs) are the most predominant group of chemical pollutants, which are generated as a result of the incomplete combustion of organic matter. The PAHs are ubiquitous in the environments and can provoke adverse health effects in the vertebrates, such as inflammation and carcinogenesis.

Skin sensitization is a human lymphocyte-mediated immune response occurring as a response to the exposure to ambient toxicants, such as PAHs. In addition, the presence of PAHs in the aquatic environments may cause increased mortality of fish and amphibians, through the PAH accumulation in the adipose tissues. The high-throughput and quantitative prediction of the toxic effects of PAHs still remains as a challenging task for the regulatory authorities.

In this work, we show a statistical investigation on the PAH toxicity outcomes in mammals and fish. The bioinformatic data relating the molecular features and toxicity endpoints were extracted and treated based on the automated workflow protocols of QSAR Toolbox. Firstly, the binary data of skin sensitization were analyzed to determine the sensitivity and specificity of PAHs. Secondly, a logistic regression method was applied to screen the skin sensitizers and non-sensitizers, associated with the physicochemical parameters of PAH molecules. Lastly, experimental and predictive datasets on the fish mortality were inter-correlated to enable the QSAR analysis.

Keywords:
toxicological data mining, OECD QSAR Toolbox, chemical toxicity prediction, statistical modeling, environmental safety

*Corresponding author, e-mail: sung@kist-europe.de

Codon usage polymorphism in SARS-2 CoV protein coding sequences

Biljana Stojanović[1], Aleksandar Veljković[2], Saša Malkov[2], Miloš Beljanski[3*], Gordana Pavlović-Lažetić[2] and Nenad Mitić[2]

[1]Mathematical Institute of the Serbian Academy of Sciences and Arts, Kneza Mihaila 36, 11000 Belgrade, Serbia
[2]Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[3]Institute for General and Physical Chemistry, Studentski trg 12, 11000 Belgrade, Serbia

Abstract
Severe acute respiratory syndrome corona virus 2 (SARS-CoV-2), which occurred in December of 2019, causes a severe acute respiratory illness, and has spread around the world. To reveal and understand genome expression mechanisms and evolution, we analyzed the codon usage (CU) pattern of SARS-CoV-2. CU refers to differences in the frequency of synonymous codons occurrence during protein translation.
Multiple sequence alignment of different types SARS CoV 2 coding sequences (ORFs) has been performed in order to reveal if there is specificity of CU for different proteins types within viral genome during pandemic. We analyzed around 66000 complete SARS-CoV-2 isolates and around 770000 coding sequences (ORFs) divided into main 12 protein groups. Protein coding sequences in each group were aligned against corresponding ORF from referent SARS-CoV-2 genome (NC_045512.2) using Clustal Omega. Aligned ORFs were analyzed for changing frequencies of G, A, T and C nucleotides at the first, second and third codon position, as well as their influence to CU.
Results obtained are illustrated on around 64000 Spike surface glycoproteins sequences.

Keywords:
SARS-CoV-2, codon usage, Spike surface glycoprotein, bioinformatics

*Corresponding author, e-mail: mbel@matf.bg.ac.rs

Myeloid derived suppressor cells-therapy attenuates experimental autoimmune encephalomyelitis and modulates gut microbiota composition

Dušan Radojević[1], Marina Bekić[2], Alisa Gruden-Movsesijan[2], Nataša Ilić[2], Saša Vasilev[2], Miroslav Dinić[1], Nataša Golić[1], Dragana Vučević[3], Miodrag Čolić[2], Sergej Tomić[2], Jelena Đokić[1]

[1]*Laboratory for Molecular Microbiology, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade*
[2]*Department for Immunology and Immunoparasitology, Institute for the Application of Nuclear Energy, University of Belgrade*
[3]*Medical Faculty of the Military Medical Academy, University of Defense in Belgrade*

Abstract

The role of gut microbiota composition in efficacy of various immune-based therapies is increasingly recognized. Thus, the aim of our study was to investigate if the efficacy of myeloid-derived suppressor cells (MDSC)-Prostaglandin E2 (PGE2) therapy for multiple sclerosis (MS) correlates with gut microbiota composition and function. MDSC generated from bone marrow cells in the presence of PGE2 were applied to spinal cord homogenate/CFA-induced experimental autoimmune encephalomyelitis (EAE) in Dark Agouti (DA) rats, an animal model of MS. MDSC-PGE2 therapy resulted in a significant attenuation of EAE symptoms over 30 days of disease monitoring. These results correlated with lower percentage of proinflammatory interferon-gamma and interleukin-17 producing cells and higher percentage of anti-inflammatory IL-4 producing cells in spinal cord and spleen. Gut microbial composition were studied using amplicon(16S rRNA)-based metagenomic analyses of fecal samples collected prior to the induction of EAE and MDSC-PGE2 therapy application, and at the peak of the disease. The induction of EAE resulted in a decrease of microbiota diversity, whereas the MDSC-PGE2 therapy preserved the diversity in EAE-induced animals. The induction of EAE in control group associated with a higher relative abundance of *Peptococcaceae,* but the lower levels of *Veillonellaceae* and different groups of *Prevotellaceae*, known to produce immunosuppressive short chain fatty acid (SCFA), and *Lactobacillus reuteri*, known for its anti-inflammatory function. In contrast, there were no changes in levels of these immunoregulatory taxa in EAE-animals treated with MDSC-PGE2 therapy. Also, SCFA producing *Ruminococcaceae,* and *Coriobacteriaceae,* known to metabolize phytoestrogens to immunosuppressive metabolites were more abundant in EAE-animals treated with MDSC-PGE2 therapy. Predicted metabolic profiling obtained by PICRUSt2 revealed that pathways involved in biosynthesis of polyamines, metabolites known to contribute to homeostasis of gastrointestinal mucosa, were enriched in MDSC-PGE2 treated animals. Considering these results, the modification of gut microbiota composition and function could further increase efficacy of MDSC-PGE-2 based therapy of autoimmune diseases.

*Corresponding author, e-mail:

Sphingosine kinase 1 inhibitors as therapeutics against cancer

Faez Iqbal Khan and Dakun Lai

School of Electronic Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610054, PR China

Abstract

Sphingosine kinase 1 (SphK1) is a promising therapeutic target against several diseases including mammary cancer. Structural folding and unfolding of SphK1 has been studied using several denaturants using CD, fluorescence spectroscopy and molecular dynamics simulations approaches. It has been found that SphK1 follows a biphasic unfolding transition (N⇌I⇌D) with an intermediate (I) state populated around 4.0 M urea concentration. The circular dichroism ($[\theta]_{222}$) and fluorescence emission spectra (λ_{max}) of SphK1 with increasing concentrations of urea were analyzed to calculate Gibbs free energy (ΔG^0) for both the transitions (N ⇌ I and I ⇌ D). A significant overlap of both the transitions obtained by two spectroscopic properties ($[\theta]_{222}$ and λ_{max}) was observed, indicating that both N ⇌ I and I ⇌ D transition follow two-step equilibrium unfolding pattern. Further, a potent lead compound using high throughput virtual screening was identified. A total 20,800 hits were identified in molecular and virtual libraries, which were reduced to 621 by several parameters of drug-likeness, lead-likeness, and PAINS. Finally, 55 compounds were selected by ADMET descriptors carried forward for molecular interaction studies with SphK1. The binding energy (DG) of three screened compounds exhibited stronger than standard drug PF-543 (–9.9 kcal/mol). Finally, it was observed that the new drug binds tightly to the catalytic site of SphK1 and remains stable during MD simulations. This study provides a significant understanding of SphK1 inhibitors that can be used in the development of potential therapeutics against breast cancer.

Keywords:

SphK1; breast cancer; molecular docking; MD simulation; MMPBSA calculations

*Corresponding author, e-mail: khanfaeziqbal@gmail.com

3d printing technologies for drug delivery: text mining of biomedical literature

Jelena Djuris*, Jelena Njegomir, Ljiljana Novovic, Sandra Cvijic and Svetlana Ibric

Department of Pharmaceutical Technology and Cosmetology, University of Belgrade – Faculty of Pharmacy, Vojvode Stepe 450, 11221 Belgrade, Serbia

Abstract

During the last decade 3d printing (3DP) technologies have revolutionized the field of biomedicine. 3DP is broadly used for tissue engineering; it allows versatility in personalization of medicines, specific drug release patterns, etc. Several 3DP technologies are currently being investigated, and there is a great need to identify materials suitable for 3DP, as well as specific delivery platforms and administration routes, in order to accelerate development of novel drug delivery systems.

The aim of this study was to utilize the text mining tools to identify research trends and predominant materials used for 3DP of drug delivery systems. Python-based, open-source Orange 3.27.1. software package was used for the text mining. Scientific articles' abstracts were retrieved from PubMed database. Once the relevant abstracts were obtained through the keyword search ("3d printing" OR "3d printed" AND "drug delivery"), text was preprocessed by removing urls, regular expressions tokenization and by using stop-words. Analysis was based on one-word terms. Word clouds were used to visualize the text corpus. Topics of research interest and correlation between the most frequently occurring words were analyzed by topic modeling and network analysis. Word enrichment was used to identify statistically significant words for a specific topic of interest, based on calculated probabilities and false discovery rates.

Text mining of 550 abstracts on 3DP in drug delivery has revealed specific drug delivery systems and administration routes, and, most importantly, the appropriate 3D printable materials (especially polymers). In terms of tissue (bone) engineering, drugs and biomolecules have been mostly formulated into scaffolds as delivery systems, followed by implants, due to a great potential for localized and targeted drug delivery. 3DP of personalized medicines is reported in a quarter of published abstracts (24.5%), with the words *tablets* and *extrusion* being recognized as statistically significant ($p<0.01$) for this topic. Hot-melt extrusion coupled with the fused deposition modeling was identified as the predominant (27%) 3DP technique. The obtained results provide valuable guidance for selection of the appropriate delivery system, 3DP technology, and printable materials based on the desired target site, duration and potential for personalization of a particular drug´s therapeutical action.

Keywords:

text mining, topic modeling, 3d printing, drug delivery, pharmaceutics, word enrichment

*Corresponding author, e-mail: jelena.djuris@pharmacy.bg.ac.rs

Optimization of Parameters for Preparing Gelatine Electrospun Microfibers

Katarina Virijević[1], Jelena Grujić[1], Mihajlo Kokanović[2], Nevena Milivojević[1], Marko N Živanović[1,2*], and Nenad Filipović[3]

[1] Institute for Information Technologies, Jovana Cvijića bb, Kragujevac, Serbia
2 BioIRC- Bioengineering Research and Development Centre, Prvoslava Stojanovića 6, Kragujevac, Serbia
3 Faculty of Engineering, University of Kragujevac, Sestre Janjjić 6, Kragujevac, Serbia

Abstract

Electrospinning nowadays is a commonly used technique in the tissue bioengineering for the promotion of tissue supportive scaffolds. Use of electrospun fibers of nano and micro dimensions for scaffold production is the key factor in creation of extracellular matrix reminding materials for cell growth. Our investigation is focused on optimization the parameters for creating the electrospun fibers derived from gelatin in slightly acidic environment. Gelatin is widely used naturally derived biodegradable material with extraordinary physicochemical and mechanical properties. We tested a numerous combinations of gelatin solutions and electrospinning parameters for obtaining the most convenient and optimal electrospun derived scaffolds for cell seeding and tissue growth in vitro. The production of these fibers is causally related to the electrospinning parameters in use, hence the importance of optimizing this process. A series of scaffold material samples were produced and microscopically analyzed. The most optimal scaffolds were seeded with MRC-5 healthy fibroblast cells prior to analyze the biocompatibility of the scaffolds. We showed that cells well adhere to gelatin scaffolds. The MTT standardized viability assay was used to investigate the possible toxicity of created material on investigated cell line.

The examined solutions were prepared by dissolving the gelatin in various volume ratios of acetic acid and water. After the preparation, solutions were stirred at room temperature. When the polymer solutions were prepared, they were placed in the 5 mL syringes with applied voltage of 20 kV between the electrodes, thus the syringe needle and aluminum collector.

The cells were maintained according to standardized procedure in incubator with humidified atmosphere supplemented with 5% CO_2 at physiological temperature of 37 °C. The both cell lines of low passages were purchased from ECACC and were cultivated in Dulbecco's Modified Eagle Medium (DMEM) (Sigma, D5796) cell culture medium supplemented with 10% fetal bovine serum (Sigma, F4135-500ML) and 1% penicillin/streptomycin (Sigma, P4333-100ML) in 75 cm^2 culture flasks. After a few passages and a confluence of about 80%, the cells were used in all in vitro experiments.

Keywords:

electrospinning, gelatin, tissue engineering

Acknowledgement:

This research is supported by the Serbian Ministry of Education, Science, and Technological Development [451-03-9/2021-14/200378 (Institute for Information Technologies, University of Kragujevac)].

*Corresponding author, e-mail: zivanovicmkg@gmail.com

Bioinformatic characterization and validation of *miR-30a-5p* and *miR-139-5p* as diagnostic and prognostic biomarkers of oral cancer

Goran Stojkovic[1,2], Ivan Jovanovic[3], Aleksandra Stankovic[3], Milovan Dimitrijevic[1,2], and Katarina Zeljic[4*]

[1] Clinic for Otorhinolaryngology and Maxillofacial Surgery, Clinical Center Serbia, Pasterova 2, 11000 Belgrade, Serbia
[2] Faculty of Medicine, University of Belgrade, Dr Subotica 8, 11000 Belgrade, Serbia
[3] „VINČA" Institute of Nuclear Sciences - National Institute of the Republic of Serbia, University of Belgrade, Mike Petrovica Alasa 12-14,11351 Vinca, Belgrade, Serbia
[3] Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia

Abstract

Oral cancer is the most prevalent type of head and neck cancer, characterized by rising incidence, high relapse occurrence and low survival. There is a high demand for the identification of novel and sensitive diagnostic and prognostic molecular biomarkers of oral cancer. MicroRNAs (miRNAs) are considered as promising candidates. Previously conducted meta-analysis on high throughput miRNA profiling in oral cancer emphasized consistent down-regulation of *miR-30a-5p* and *miR-139-5p*. We aimed to validate these findings in oral cancer clinical samples and to test the diagnostic and prognostic potential of these miRNAs. Bioinformatical prediction of investigated miRNAs target genes was executed through the intersection of multiple miRNA target prediction algorithms. Enrichment analysis of miRNA target genes was performed using miRPathDB V2.0 software on GO:BP database. RNA was isolated from oral cancer and adjacent non-cancerous tissue from 30 patients by miRVana kit. cDNA was synthesized by TaqMan microRNA reverse transcription kit. TaqMan gene expression kit was used for relative quantification of miRNAs normalized to *RNU6B*. Relative expression was calculated by comparative Ct method. The diagnostic potential was estimated by the ROC curve analysis. Expression of *miR-30a-5p* and *miR-139-5p* was dichotomized as high or low by the median. Association with clinicopathological features and miRNAs expression was calculated by χ^2 test.

Bioinformatical analysis demonstrated both enrichment of *miR-30a-5p* and *miR-139-5p* targets in biological processes associated with cancer pathology and complementarity in regulation of these processes thus ensuring non-redundant regulation. Both miRNA were significantly down-regulated in oral cancer compared to non-cancerous tissue (Wilcoxon test: p=0.0004, p=0.001, respectively). According to the ROC curve analysis, both miRNAs might be used as potential tools for discrimination between cancerous and non-cancerous tissue (*miR-30a-5p* AUC=0.677, p=0.019; *miR-139-5p* AUC=0.656, p=0.038). There was a significant correlation between the expression of *miR-30a-5p* and *miR-139-5p* in cancerous tissue (Spearman rank test r=0.901, p<0.0001). None of the analyzed miRNAs was associated with stage, nodal status, tumour size and overall survival (p>0.05) which indicates that *miR-30a-5p* and *miR-139-5p* can't be used as prognostic biomarkers of oral cancer.

Our results suggest that *miR-30a-5p* and *miR-139-5p* might be used as good diagnostic biomarkers for discrimination between oral cancer and non-cancerous tissue.

Keywords:
oral cancer, microRNA, *miR-30a-5p*, *miR-139-5p*, expression.

*Corresponding author, e-mail: katarina.zeljic@bio.bg.ac.rs

Prediction of short-term success of electrical cardioversion

Lazar Vasović[1*] and Ana Jakovljević[1]

[1] *Faculty of Mathematics, University of Belgrade, Studentski trg 16,*
11000 Belgrade, Serbia

Abstract

Electrical cardioversion is a medical technique that uses synchronized electrical shocks to restore normal heart rhythm in people with persistent arrhythmia. This kind of heart rate problem is usually associated with a disease called atrial fibrillation.

The aim of this paper was to create a classification model that accurately predicts whether the procedure will be successful in the short term (immediate outcome), based on data on the clinical picture, other indications, and drug therapy prescribed to patients undergoing it. Dataset, consisting of 147 unique instances, was obtained from the Pacemaker Center of the Clinical Center of Serbia and pertains to patients with electrical cardioversion performed from 2014 to 2019. It is noticeably imbalanced with concern to target class; Successful procedures were marked as class True (130 – 88.4%) and unsuccessful as class False (17 – 11.6%).

The focus was on Bayesian networks, a well-known probabilistic graphical model. They, however, proved inferior to other methods of classification and machine learning, such as the random forest classifier and artificial neural networks. Fitted models were compared by their accuracy, sensitivity (recall), specificity, and other relevant metrics. Extra attention was given to data preprocessing and exploratory analysis, as well as predictor (feature) importance.

Experiments included one Bayesian network structure with fair predictive values and some other similarly successful models. A voting ensemble made of multilayer perceptron (sklearn MLPClassifier) and complement naïve Bayes (sklearn ComplementNB) turned out the best, with full 100% specificity, recall of the important class False, while maintaining a relatively high F_1 score of 63% on the same class. Accuracy was 87%, while balanced accuracy was 92%. Precision on class False was not that good 46%, but it was still the best compromise.

Results also give insight into predictor importance, such as that extracted from the decision tree classifier, which marked patient age, heart rate, and total duration of the indicated heart disease as the most significant ones. Other methods of statistical analysis and machine learning were also used to identify important features, including a dendrogram generated by clustering attributes based on their correlations as the similarity measure.

Keywords:
cardioversion, atrial fibrillation (AF), voting ensemble, Bayesian network

*Corresponding author, e-mail: mi16099@alas.matf.bg.ac.rs

De novo assembly and annotation of the marbled crayfish and noble crayfish hepatopancreas transcriptomes

Ljudevit Luka Boštjančić[1,*], Caterina Francesconi[2], Christelle Rutz[3], Lucien Hoffbeck[3], Laetitia Poidevin[3], Arnaud Kress[3], Japo Jussila[4], Jenny Makonnen[4,5], Feldmeyer Barbara[6], Miklós Bálint[1], Odile Lecompte[3], Kathrin Theissinger[1,2]

[1]LOEWE Centre for Translational Biodiversity Genomics (LOEWE-TBG), Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Georg-Voigt-Str. 14-16, 60325 Frankfurt am Main, Germany
[2]Institute for Environmental Sciences, University of Koblenz-Landau, Fortstrasse 7, 76829 Landau, Germany
[3]Department of Computer Science, ICube, UMR 7357, University of Strasbourg, CNRS, Centre de Recherche en Biomédecine de Strasbourg, Rue Eugène Boeckel 1, 67000 Strasbourg, France
[4]Department of Environmental and Biological Sciences, University of Eastern Finland, P.O. Box 1627, 70210 Kuopio, Suomi-Finland
[5]Present address: BioSafe - Biological Safety Solutions, Microkatu 1, 70210 Kuopio, Finland
[6]Molecular Ecology, Senckenberg Biodiversity and Climate Research Centre, Georg-Voigt-Str. 14-16, 60325 Frankfurt am Main, Germany

Abstract

Freshwater crayfish represent keystone species of the freshwater ecosystems. With the introduction of invasive North American crayfish species and their pathogen *Aphanomyces astaci* a significant decline in the size and abundance of European freshwater crayfish populations was observed. Marbled crayfish is an invasive species exhibiting a high resistance to the *A. astaci* challenge. On the other side, noble crayfish, native European species is highly susceptible to the *A. astaci* infection. Exploring molecular mechanisms of resistance on the transcriptome level represents the next milestone for the crayfish immunology. In total, hepatopancreas tissues was isolated form 25 noble crayfish and 30 marbled crayfish. This was followed by RNA sequencing on the Illumina NovaSeq 6000 platform. Quality control of raw reads was conducted with FastQC, followed by the adaptor and quality trimming conducted in Trimmomatic. *De novo* transcriptomes were assembled with Trinity, resulting in 109608 (N_{50} = 1360) and 254336 (N_{50} = 1082) Trinity genes for noble crayfish and marbled crayfish, respectively. BUSCO assembly scores placed these assemblies as the most complete for freshwater crayfish at 93,30% completeness for noble crayfish and 93,98% completeness for marbled crayfish. Assembled transcripts were annotated using the dammit! pipeline and assigned to KEGG pathways. Respective transcriptomes and raw datasets may be reused as the reference transcriptome assemblies for future gene expression studies.

Keywords:
Freshwater crayfish, *Astacus astacus*, *Procambarus virginalis*, crayfish plague, RNA sequencing

*Corresponding author, e-mail: luka.bostjancic@senckenberg.de

Expression analysis of *kcs* genes in sunflower under drought stress

Mahmood-ur-Rahman, Parwsha Zaib, Munazza Ijaz, Roshina Shahzadi

Department of Bioinformatics and Biotechnology, Government College University, Faisalabad, Pakistan

Abstract

Drought stress is considered as the main abiotic factor which badly affects growth of sunflower plant. Several studies have been carried out to understand the mechanism of drought stress tolerance in plants. *KCS* genes are responsible for wax biosynthesis and reported to be involvedin drought stress tolerance. In this study, comparative genomics and expression profiling of *KCS* genes was done to understand their role in stress mechanism. Phylogenetic analysis revealed that KCS genes were divided into six distinct clades which was further confirmed by Synteny analysis and concluded that *KCS* genes in both species share the same evolutionary origin. Further, they were amplified in sunflower by using gene specific primers. Five genes, i.e. *KCS2*, *KCS4*, *KCS5*, *KCS10* and *KCS18* were successfully amplified in sunflower. Then, sunflower plants were subjected to drought stress and expression profiling of amplified *KCS* genes was carried out by Real Time PCR. All the five genes were up-regulated under drought showing theirrole in stress conditions; however, the expression level of each gene was varied. Maximum relative expression was found for *KCS4* gene in T1, i.e. 19 fold as compared to control. Total chlorophyll contents were decreased under drought stress while antioxidants like catalase, peroxidase, superoxide dismutase and proline were increased. This study concluded that *KCS* genes have role in drought stress tolerance and their expression is significantly up-regulated under stress conditions.

Keywords:

Sunflower, wax biosynthesis, drought, gene expression profiling, *KCS* genes

*Corresponding author, e-mail: mahmoodansari@gcuf.edu.pk

Differential gene expression analysis of heterotic groups' maize inbred lines under optimal conditions led to the identification of specific gene regulation under low-temperature

Manja Božić[1*], Ana Nikolić[1], Dragana Dudić[3], Dragana Ignjatović-Micić[1], Jelena Samardžić[2]
Nenad Delić[1], Bojana Banović Đeri[2]

[1] Maize Research Institute „Zemun Polje", Slobodana Bajića 1, 11085 Belgrade, Serbia
2 Institute of Molecular Genetics and Genetic Engineering, Vojvode Stepe 444a, 11042 Belgrade, Serbia
3 Faculty of Informatics, University Union-Nikola Tesla, Cara Dušana 62-64, 11158 Belgrade, Serbia

Abstract
Finding new ways of improving crop quality, yield potential and abiotic stress tolerance are some of the most important pursuits in crop production today. As one of the biggest causes of yield and productivity reduction is climate change, specifically increasing temperatures and drought during the summer, a large number of strategies is focussed on lessening their negative effects. Cropping pattern changes include earlier sowing (early spring), when the temperatures are lower, as one of the most promising escape strategies for avoiding high summer temperatures. Thus, development of cold tolerant maize lines became an important goal. Comparative analysis of 46 maize inbred lines belonging to two different genetic backgrounds, one predominantly cold tolerante (marked as Non-Lancaster) and the other predominantly cold sensitive (marked as Lancaster) in the field, was done by whole transriptome sequencing and differential gene expression (DGE) analysis. Plants were grown under optimal, greenhouse conditions and sampled after completing the V4 growth stage. Total RNA isolated from leaves of three plants per inbred line was used for cDNA library preparation by Illumina TruSeq Stranded RNA LT kit. Pair-end sequencing was performed on MiSeq Illumina sequencer using MiSeq Reagent kit, v2 (2 x 150bp). Data manipulation and analysis was performed using a custom-made bioinformatics pipeline that included high throughput sequence data quality control (using FastQC), removal of low quality reads (using Trimmomatic tool, version 0.32), transcriptome assembly and mapping (using Cufflinks, version 2.2.1), expression quantification (using CuffDiff) and DGE analysis (using BLAST2GO and GO analysis Toolkit and Database for Agricultural Community, agriGO v2).
DGE analysis revealed 77 differentially expressed genes (DEGs) between the Lancaster and the Non-Lancaster group, 21 of which were statistically supported for differential expression between the two groups and annotated as involved in abiotic stress responses in maize and other plant species. To test DEGs response to cold stress expression of a subset of seven DEGs in eight inbred lines (4 belonging to Lancaster and 4 belonging to Non-Lancaster genetic background) was analyzed under 24^h long exposure to low temperatures (6/4° C, 12^h photoperiod), with sampling being done 6^h and 24^h after beginning of the treatment, as well as after 48^h of recovery. Six DEGs showed different expression regulation dependent on cold exposure duration and genetic background. These findings imply differently regulated processes between the analysed Lancaster and Non-Lancaster inbred lines, contributing to their different cold response and adaptation, and will be further used for the development of cold tolerant hybrids.

Key words:
Transcriptomics, NGS, DEGs, maize, cold tolerance

*Corresponding author, e-mail: mbozic@mrizp.rs

Comparative *De Novo* Transcriptomic Analysis of Photosynthetically Active and Non-Photosynthetically Active Tissues of Variegated *Pelargonium zonale* Leaves

Marija Vidović[1*], Bojana Banović Djeri[1], Jelena Samardžić[1]

[1] *Institute of Molecular Genetics and Genetic Engineering, Laboratory for Plant Molecular Biology, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia*

Abstract

Variegated *Pelargonium zonale* leaves have proven to be an excellent model system to examine source–sink interactions within the same organ providing the equal microenvironment conditions, unlike common shoot/root relation studies. Photosynthetically non-active (W) mesophyll cells contain smaller plastids lacking thylakoid membranes or starch granules, and exhibit no peroxisomes in comparison to photosynthetically active (G) cells. With the aim of gaining a deeper insight into molecular phenotype of W leaf tissue, particularly the one related to photosynthetic-dependent H_2O_2 metabolism, transcriptomes of these two metabolically contrasted tissues were compared.

High-quality total RNA from W and G leaf tissues was extracted according to our previously optimised protocol. Highly purified cDNA libraries were synthesized and sequenced on an Illumina platform. The ambiguous nucleotides, adapter sequences, and low-quality sequences were trimmed and the read quality was checked before and after the trimming. In total, 39763284 (with Q30=94.3%) and 42062153 (with Q30=94.0%) clean reads were obtained in G and W total RNA samples, respectively, and used to perform transcriptome assembly by Trinity software. After removing the redundancy, via Corset software, 139811 transcripts with 139575 unigenes were annotated through comparison with seven commonly used databases (NCBI non-redundant protein and nucleotide sequences; PFAM; Clusters of Orthologous Groups of proteins, Swiss-Prot, KEGG, GO).

Analysis of differentially expressed genes was performed using DESeq2 R package and revealed 4668 up-regulated genes and 6689 down-regulated genes in G tissue compared with W one. Among the up-regulated genes in G tissue, the majority was associated with cytoskeleton, photosynthetic processes, plastids, thylakoids and transport, while in W tissue up-regulated genes were mainly found to encode enzymes with ATPase activity, carbohydrate absorption and digestion, callose, pectin and linoleic acid metabolism. Moreover, a significant difference between these two tissues differing in H_2O_2 generation rate was observed in the expression level of genes involved in H_2O_2 scavenging. Enzymatic constituents of the ascorbate-glutathione cycle and glutathione-S-transferase were up-regulated in W tissue, while catalase, glutathione-peroxidases and three Class III peroxidases were all up-regulated in G tissue. The obtained transcriptome results were correlated with previously revealed morphological, biochemical, and molecular characteristics of these two tissues.

Keywords:

antioxidative metabolism, differential gene expression analysis, H_2O_2 scavenging, photosynthesis, source/sink metabolism, variegated plants.

Acknowledgements:

This work was funded by the Ministry of Education, Science and Technological Development, Republic of Serbia (Contract No. 451-03-9/2021-14/ 200042), Bilateral project (no. 451-03-01963/2017-09/09) and Centre for Green Technologies, University of Belgrade.

*Corresponding author, e-mail: mvidovic@imgge.bg.ac.rs

In silico study of some tetra- and penta-coordinated gold(III) complexes as potential inhibitors of SARS-CoV-2 main protease

Marko Antonijević[1*], Ana Kesić[1], Dejan Milenković[1], Jelena Đorović Jovanović[1] and Zoran Marković[1]

[1] *University of Kragujevac, Institute for Information Technologies, Jovana Cvijica bb, 34000 Kragujevac, Serbia*

Abstract

During the last 20 years, much interest has been focused on some gold(III) complexes, due to their stability under physiological-like conditions. Gold(III) complexes have a significant effect on enzyme inhibition due to their strong sulfur binding affinity towards various sulfur-containing enzymes such as thioredoxin reductase, glutathione reductase, and cysteine protease. With the advent of SARS-CoV-2 viral infection, there has been a need to find inhibitors that will prevent the virus from acting. Recent research has shown that compounds that have certain functional groups in their structure are effective in inhibiting the main protease of this virus. In this paper, the inhibitor efficiency of the gold(III) complexes ($[Au(DPP)Cl_2]^+$ (**C1**) and $[Au(DMP)Cl_3]$ (**C2**), where DPP=4,7-diphenyl-1,10-phenanthroline and DMP=2,9-dimethyl-1,10-phenanthroline), as well as FDA approved drugs, cinanserin and chloroquine towards the main protease of SARS-CoV2 (M^{pro}) was estimated using the molecular docking simulations. The binding affinity of investigated compounds was examined by the AutoDock 4.2 software. The ligands were prepared for docking by optimization of their geometries by density functional theory (DFT) employing M06-2X functional in combination with the 6-311G(d,p) basis set for C, N, S, Cl, and H, and LAN2DZ basis set for Au. The native bound ligand (N3) was extracted from M^{pro} and binding pocket analysis was performed by the AutoGridFR program. Re-docking was performed with the investigated compounds to generate the same docking pose as found in the co-crystallized form of M^{pro}. Analysis by AGFR showed that the investigated compounds bind in the active site of M^{pro}. The obtained results indicate that the square-planar **C1** shows better inhibitory activity compared to cinanserin and chloroquine. The binding free energy of **C1** is significantly higher than that for FDA drugs, with values of -38.4, -31.8, and -31.2 kJ mol^{-1}, respectively. The obtained results revealed that **C1** and **C2** bind at the same binding pockets to M^{pro} as well as FDA drugs by weak non-covalent interactions. The most prominent interactions are hydrogen bonds, alkyl-π, and π-π interactions. The preliminary results suggest that gold(III) complexes showed good binding affinity against M^{pro}, as evident from the free binding energy (ΔG_{bind} in kJ/mol).

Keywords:

gold(III) complexes, DFT, molecular docking, SARS-CoV2

Acknowledgements:

Authors acknowledge the Ministry of Education, Science and Technological Development of the Republic of Serbia for the support through Contracts No. 451-03-09/2021-14/200378.

*Corresponding author, e-mail: mantonijevic@uni.kg.ac.rs

Bioinformatics pipeline for genotyping and genotype - phenotype association study in maize (*Zea mays* L.)

Marko Mladenović[1*], Nikola Grčić[1], Dragana Dudić[4], Ana Nikolić[1], Manja Božić[1], Nenad Delić[1], Slaven Prodanović[3], Bojana Banović Đeri[2]

1 Maize Research Institute "Zemun Polje", Slobodana Bajića 1, 11085 Belgrade, Serbia
2 Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
3 Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11080 Belgrade, Serbia
4 Faculty of Informatics, University Union-Nikola Tesla, Cara Dušana 62-64, 11158 Belgrade, Serbia

Abstract
Multidisciplinary research is today commonly used in plant breeding for improving important agronomic traits. High throughput genotyping technologies and genotype – phenotype association studies as widely used for improving breeding programs, depend on bioinformatics analysis for extracting information from the gathered data. In this research, among plethora of widely used bioinformatics approaches, the custom made one was chosen, based on the current recommendations in the field.

The material includes a set of 46 maize inbred lines commonly used in maize breeding programs. Phenotyping was done for thirteen important quantitative agronomic traits in 8 environments during two years (2018 and 2019). For the purpose of genotyping, plants of all inbred lines were grown under optimal conditions and sampled after completing the V4 growth stage. Total RNA was isolated from the third leaf of three plants per inbred line and used for cDNA preparation by Illumina TruSeq Stranded RNA LT kit. Pair-end RNA-Seq based on Next Generation Sequencing methodology was performed on MiSeq Illumina sequencer using MiSeq Reagent kit, v2 (2 x 150bp). Raw sequencing data of maize leaves' transcriptionally active genome regions at the moment of sampling were used for identification of single nucleotide polymorphisms (SNPs) in each of 46 inbred lines.

Bioinformatics pipeline for data manipulation and analysis was custom made and included FastQC (for quality control (QC) of raw data), Trimmomatic tool v0.32 (for adapter and contaminants removal, as well as for the removal of regions with QC below 30), TopHat (insert size 130, standard deviation 50, maximum intron size 100.000 – for mapping filtered reads onto the B73 maize reference genome v3.0), Cufflinks v2.2.1 (for reads assembly), Cuffmerge (for the final transcriptome assembly) and an intersection output of two independent SNPs calling tools FreeBayes and BCFtools (to minimize false positive results). With the aim to find SNP markers which show strongly statistically supported relationship with favorable values of investigated quantitative traits, genotype - phenotype association analysis was conducted. It was performed using two approaches – one relying on the TASSEL software, widely used in agronomics and the other based on machine learning software like WEKA, rarely used in agronomics. The results of two approaches were compared and discussed.

Keywords:
maize, bioinformatics, genotyping, RNA-Seq, genotype-phenotype association

*Corresponding author, e-mail: mmladenovic@mrizp.rs

A new model for prediction of bacterial chromosome replication origin

Marko Stojićević[1*], Miloš Beljanski[2], Nenad Mitić[1]

[1] Faculty of Mathematics, University of Belgrade, Studentski trg 16,
11000 Belgrade, Serbia
[2] Institute of General and Physical Chemistry, Studentski trg 12,
11000 Belgrade, Serbia

Abstract

DNA replication is one of the most basic and important biological processes. Bacteria usually contains circular DNA chromosome with specific locus designated as origin of replication - oriC. For better understanding of the mechanism of the DNA replication's initiation, it is important to identify the position of replication origin. OriC positions are experimentally confirmed only for few bacteria. Additionally, the prediction of OriC is possible through mathematical prediction methods and computer algorithms. There are several methods for OriC identification that are commonly used, but none of those methods give the correct positions. The cumulative GC (CGC) skew method uses the sliding window strategy, where the CGC skew plot peaks should correspond to terminus/origin position. However, calculating CGC skew, the resulting values are not in reach of experimentally confirmed OriC position. The aim of this work was to upscale the CGC skew method, by adding additional calculations and analysis, in order to get more precise OriC positions. As a basis for model 37 circular bacterial genomes with experimentally confirmed OriC position (provided by the NCBI database) were analyzed. For every genome, a CGC skew analysis and plot has been driven.

The results show that the function's peak is not close enough to the confirmed OriC position. With further analysis, it can be concluded that the experimentally confirmed OriC is typically inside the sliding window (or in the first following window) that contains the first function's inflection point following the function's maximum point. A genome sequence, that consists of nucleotides from three sliding windows (one with the function's inflection point, one before, and one after), was searched for the repeating sequences. Minimal length of nine nucleotides was taken as an average length of statistically significant repeats for all of the different genomes that were analyzed. Due to different strain positioning, different types of repeats were analyzed – direct/inverse non-complementary and direct/inverse complementary repeats. From the obtained results it can be concluded that this new model gives promising results in comparison to standard CGC skew, in terms of more precise OriC positions.

Keywords:

bioinformatics, bacterial chromosome, replication origin, cumulative GC skew

*Corresponding author, e-mail: mstojicevic@gmail.com

Repurposing of antiparasitic drugs for Candidate SARS-CoV-2 Main Protease Inhibitors by combined *in silico* Method

Milan Sencanski*, Jelena Milicevic, Vladimir Perovic and Sanja Glisic

VINCA Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade, Mike Petrovića Alasa 12-14, 11351 Belgrade, Serbia

Abstract

The SARS-CoV-2 outbreak that is spreading rapidly around the world requires urgently effective treatments. Therefore, *in silico* drug repurposing represents a powerful strategy to enable the acceleration of the identification of drug candidates with already known safety profiles. The SARS-CoV-2 main protease is essential for viral replication and an attractive drug target. This study used the virtual screening protocol with both long-range and short-range interactions to select candidate SARS-CoV-2 main protease inhibitors. The Informational spectrum method developed for small molecules was first applied for searching the Drugbank database of antiparasitic agents and further followed by molecular docking. After *in silico* screening of drug space, we propose several drugs as potential SARS-CoV-2 main protease inhibitors for further experimental testing.

Keywords:

SARS-CoV-2, main protease, virtual screening, drug repurposing, antiparasitics

*Corresponding author, e-mail: sencanski@vin.bg.ac.rs

Prediction of GO terms for IDPs based on highly connected components in PPI networks

Milana Grbić[1*], Branislava Gemović[2], Radoslav Davidović[2], Aleksandar Kartelj[3], Dragan Matić[1]

[1] Faculty of Natural Sciences and Mathematics, University of Banja Luka, Mladena Stojanovića 2, 78000 Banja Luka, Bosnia and Herzegovina
[2] "VINČA" Institute of Nuclear Sciences - National Institute of the Republic of Serbia, University of Belgrade 11000 Belgrade, Serbia
[3] Faculty of Mathematics, University of Belgrade, Studentski trg 16,
11000 Belgrade, Serbia

Abstract

Partitioning large biological networks can help biologists to retrieve new information for particular biological structures. In literature, various methods for partitioning and clustering biological networks have been proposed. The aim of such a network partitioning is to retrieve smaller structures which are easier to analyse, but still containing important information about relations between the network elements.

Highly connected deletion problem is one of such network partitioning, with the aim to partition a network into highly connected components (hcd components) by deleting minimum number of edges. A network component with n nodes is a hcd component if the degree of every vertex is larger than n/2. For the purpose of this research, we used a specially constructed local search based heuristic approach to identify hcd components.

Dealing with protein-protein interaction (PPI) networks, it has been noticed that proteins from the same hcd component in a network have same Gene Ontology (GO) annotations. Based on that, we proposed a new method for prediction of GO annotations, which consists of the following steps:

(a) starting PPI network is partitioned to hcd components;
(b) the obtained hcd components are expanded by proteins which became singletons in the partition set;
(c) the newly formed extended hcd components are the subject of further enrichment analysis in DiNGO tool, which returns a list of existing GO terms for proteins from the considered extended component;
(d) after propagation through GO hierarchy, the extended list of GO is obtained;
(e) each protein from the extended hcd component is annotated by a number of GO terms obtained from the previous step;

The proposed method is tested on the data from CAFA-3 challenge. Comparing the F1-measure of the obtained results, a combination of parameters (type of extension, cutoff for enrichment analysis and maximum number of GO terms) with the best performances is selected for the further usage. The method with the selected parameters was further applied on a class of Intrinsically Disordered Proteins (IDP). Preliminary results indicate that this method can be useful for proposing new GO terms for IDP proteins.

Keywords:
hcd components, GO terms, enrichment analysis, protein function annotation

*Corresponding author, e-mail: milana.grbic@pmf.unibl.org

Brevibacillus laterosporus supplementation diet modulates honey bee microbiome

Milka Malešević[1], Slađan Rašić[2], Violeta Santrač[3], Milan Kojić[1], Nemanja Stanisavljević[1]*

[1]*Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia*
[2]*Educons University, Vojvode Putnika 85-87, 21208 Sremska Kamenica, Serbia* [3]*Veterinary Institute of Republic of Srpska "Dr. Vaso Butozan", Branka Radičevića 18, 78000 Banja Luka, Republic of Srpska*

Abstract
Honey bees (*Apis mellifera*) are facing multiple stressors affecting their lifespan, health and productivity. Among them, bacterial and fungal pathogens *Paenibacillus larvae*, *Melissococcus pluton*, *Ascosphera apis* and *Nosema ceranae* play a major impact on honey bees colonies. Thus, developing alternative prophylactic and curative strategies are urgently needed. The use of probiotic bacteria in honey bee supplemental feeding is therefore promising to treat or prevent diseases. *Brevibacillus laterosporus*, Gram-positive endospore forming bacilli, is recognised as one of the promising antibacterial and antifungal agents producer.

The aim of this study was to examine the short-therm effects of *B. laterosporus* supplemented diet on worker honey bee microbiome.

Dry spores of *B. laterosporus* strain BGSP11 have been administrated through a sugar syrup diet to ten colonies and a representative specimen of worker honey bees was taken before the start of the treatment and immediately after the syrup was consumed. The microbial diversity was assessed before and after the treatment using Illumina MiSeq sequencing platforms (ID Genomics service, Seattle, WA, USA). 16s rRNA sequencing for bacterial community profiling and fungal Internally Transcribes Spacer for mycological taxa profiling were used. The next-generation microbiome bioinformatics platform QIIME2 v 2021.4 was used for filtering and denoising obtained sequences, calculation of diversity metrics and taxonomy assignment. The feature classifier was trained using the Greengenes v 13_8 for bacterial taxa and fungal UNITE database v 8.3. The results obtained in this study indicated statisticaly significant alfa diversity between control and experimental group honey bee microbiota composition. The diversity abundance was higher in control comparing to the group treated with *B. laterosporus* strain BGSP11 spores. There was no significant diference in Bray-Curtis distance among two groups of analysed samples. Regarding to mycological abundance, composition was completely different between two groups; control group had *Claviceps* as predominant genus, while in treated group of honey bee microbiome *Metschnikowia* genus was prevalent, indicating that the presence of fungal pathogens in treated group is highly diminished.

Keywords:
Brevibacillus laterosporus, honeybee microbiome, metagenomics, 16S rRNA sequencing, Internally Transcribes Spacer sequencing

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia [451-03-9/2021-14/200042].

*Corresponding author, e-mail: milkam@imgge.bg.ac.rs

Identification of differentially expressed genes in SARS-Cov-2 infected cells using Bayesian network models

Nevena Ćirić[1*], Aleksandar Veljković[1]

[1] *Faculty of Mathematics, University of Belgrade, Studentski trg 16,*
11000 Belgrade, Serbia

Abstract

The current outbreak of infectious disease caused by a novel coronavirus, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been affecting millions of people and has caused devastating mortality worldwide. In this regard, development of drugs, vaccines and treatments addressing the SARS-CoV-2 infection have become a major focus. The identification of differentially expressed genes due to SARS-CoV-2 infection may provide valuable information about the underlying biology of the disease. It can give an insight into molecular mechanisms of disease by indicating the signaling pathways altered during the infection and finding key molecular players that can be targeted.

Network analysis is the most convenient method for representation of a functionally related set of genes and detection of changes in their expression. This study builds upon the Bayesian network model and coexpression network analysis applied to identification of differentially expressed genes in SARS-CoV-2 infected cells. Weighted gene coexpression network analysis (WGCNA) is used to group related genes into gene modules based on their coexpression patterns in non-infected cells. For each gene module, WGCNA computes one eigengene – weighted average of the expression of all the genes in that module, whereby weights are determined so that loss in the biological information is minimized. These eigengenes are used to train a Bayesian network in which nodes (random variables) represent gene modules and directed edges represent the conditional dependencies between corresponding gene modules. Besides random variables that model the expression value of each eigengene, the network has one additional binary variable which models type of sample – infected or non-infected. According to the Markov property of the Bayesian networks, the parents of that node are the modules most related to, thus they should be enriched with genes that are associated with the disease.

This task has been performed elsewhere to specific types of cells and using different network analysis methods - gene ontology, pathway enrichment analysis and functional protein network construction. The results obtained may serve as a doubled evidence of an important finding.

Keywords:
SARS-CoV-2, gene expression, Bayesian networks, coexpression network

*Corresponding author, e-mail: nevena_ciric@matf.bg.ac.rs

Applications of neural network models in predicting enzyme function based on the EC nomenclature

Nevena Ćirić[1*], Mladen Nikolić[1], Jovana Kovačević[1]

[1] *Faculty of Mathematics, University of Belgrade, Studentski trg 16,*
11000 Belgrade, Serbia

Abstract

The Enzyme Commission Number (EC Number) represents a numerical classification scheme for enzymes that uses 4 numbers to describe the functions of enzymes based on the chemical reactions they catalyze. Each EC number corresponds to an enzyme family of a specific type, which are hierarchically organized on 4 levels. The classical approach to the problem of enzyme function prediction involves the selection of some chemical, physical and structural properties of enzymes based on which the so-called functional annotation transfer is used for annotation association. It includes transferring annotation from enzymes with similar properties, whose functions have been experimentally determined, to the enzyme whose function is being predicted. A more advanced approach consists of the application of neural networks - a machine learning model that is able to construct new properties based on the selected dataset, which can be more informative for predicting enzyme function.

In this study, we applied an approach that combines two types of local structural properties of enzymes and two types of neural networks that build new properties based on ones given and predict the enzyme function. Enzyme sequences were used as local information on enzyme structure, and for the local information on enzyme function, we used functional domains. The neural network model consists of two parts – fully connected and recurrent, whose inputs are, respectively, functional domains and enzyme sequences. Four models were constructed for the same neural architecture combining the use of different techniques for regularization and improvement of learning process. Also, we defined loss function that is hierarchy-related and penalizes differently depending of the distance of true and predicted values in the label hierarchy.

In predicting the input enzyme's main class all four models demonstrated stable performance and scored both globally and per-class accuracy 98.4%, while in predicting its subclass the accuracy was 97.52%. These results are not directly comparable with results of other published methods since we used a different set of enzyme properties and different versions of EC nomenclature. Nevertheless, our results are found to be of very similar order of magnitude, if not better, considering that a smaller set of properties was used.

Keywords:

enzyme function, EC number, machine learning model, neural network

*Corresponding author, e-mail: nevena_ciric@matf.bg.ac.rs

Large scale mitochondrial DNA analysis of European Honey bee (*Apis mellifera*) populations from the Balkans, population genetics and phylogeographic perspective

Pavle Erić[1 *], Aleksandra Patenković[1], Katarina Erić[1], Vanja Tanasić[3], Milica Mihajlović[3], Marija Tanasković[1], Ljubiša Stanisavljević[2], and Slobodan Davidović[1]

[1]Department of Genetics of Populations and Ecogenotoxicology, Institute for Biological Research "Siniša Stanković" – National Institute of the Republic of Serbia, University of Belgrade, Bulevar Despota Stefana 142, 11060 Belgrade, Serbia.
[2]Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia.
[3]Center for Forensic and Applied Molecular Genetics, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia.

Abstract

Local populations of *Apis mellifera* are rapidly changing, with the gene pool of autochthonous breeds being depleted by beekeepers through the import of foreign queens not adapted to the local environment. To study their genetic structure and phylogenetic relationships, we gathered a large dataset from the Balkans and surrounding countries.

Our sample consisted of 246 honeybee colonies collected from 47 apiaries and 24 feral colonies divided into four subpopulations from southern Serbia, five from Vojvodina, and two from Belgrade. To evaluate genetic diversity patterns, we sequenced the mitochondrial *tRNAleu-cox2* intergenic region. We compared our data to other published data on *A.mellifera COI-COII* intergenic region variability in the Balkans and neighboring countries. We pulled 1512 sequences from the NCBI GenBank, originating from 15 different populations.

The 1782 mitochondrial sequences were grouped into 31 haplotypes, with two newly described haplotypes from our sample. All haplotypes belonged to the eastern Mediterranean C lineage. The most frequent haplotype was C2d, characteristic for A.m.macedonica, followed by C2c and C1a characteristic for A.m.carnica and A.m.ligustica respectively. In our samples 9 haplotypes were observed, with the C2d being the most common and widespread as it was detected in all 11 groups, followed by C2e that was detected in all but one group. C2c and C1a were a little less common than in the total sample but very widespread as they were present in seven groups.

When the Macedonia, Ukraine, and Belgrade honeybee populations which consisted of a single haplotype, were excluded, the haplotype diversity ranged from 0.0998 to 0.7477, nucleotide diversity ranged from the lowest value of 0.000114 to 0.003731. The mean number of pairwise differences for populations that had more than one haplotype ranged from 0.060577 to 2.

MDS plot constructed on pairwise F_{ST} values shows significant geographical stratification, with our subpopulations being grouped together. Vojvodina being placed closer to Romania and Hungary datasets, while southern Serbia is closer to Bulgaria and Montenegro. Interestingly, our samples are not closely grouped with the Serbian dataset from the GenBank which indicates that honeybee populations are changing rapidly.

Keywords:

Apis mellifera, genetic diversity, Subspecies, Mitochondrial DNA, *COI-COII* intergenic region

*Corresponding author, e-mail: pavle.eric@ibiss.bg.ac.rs

Genetic diversity analysis of microsatellites and mitochondrial *Cytb* gene, relatedness estimates and *Cytb* phylogeography of protected Griffon vulture species from Serbia

Slobodan Davidović[1], Mihailo Jelić[2], Saša Marinković[3], Mila Kukobat[2], Milica Mihajlović[4], Vanja Tanasić[4], Irena Hribšek[5], Goran Sušić[6], Milan Dragićević[7], Marija Tanasković[1] and, Marina Stamenković-Radak[2]

[1]*Department of Genetics of Populations and Ecogenotoxicology, Institute for Biological Research "Siniša Stanković" – National Institute of the Republic of Serbia, University of Belgrade, Bulevar Despota Stefana 142, 11060 Belgrade, Serbia.*
[2]*Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia.*
[3]*Department of Ecology, Institute for Biological Research "Siniša Stanković" – National Institute of Republic of Serbia, University of Belgrade, Bulevar Despota Stefana 142, 11060 Belgrade, Serbia.*
[4]*Center for Forensic and Applied Molecular Genetics, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia.*
[5]*Birds of Prey Protection Foundation, Bulevar Despota Stefana 142, 11060 Belgrade, Serbia.*
[6]*Ornithological Station Rijeka, Croatian Academy of Sciences and Arts, Ružićeva 5/2, 51000 Rijeka, Croatia.*
[7]*Department of Plant Physiology, Institute for Biological Research "Siniša Stanković" – National Institute of the Republic of Serbia, University of Belgrade, Bulevar Despota Stefana 142, 11060 Belgrade, Serbia*

Abstract

Once a widespread species across the region of Southeast Europe, the Griffon vulture is now confined to small and isolated populations across the Balkan Peninsula. The population from Serbia represents its biggest and most viable population that can serve as an important reservoir of genetic diversity from which the birds can be used for the region's reintroduction programs. The available genetic data for this valuable population are scarce and it is necessary to assess its genetic diversity and inbreeding level if the population is going to be used for restocking and reintroduction.

To assess the genetic diversity we used microsatellite markers from ten loci and mitochondrial *Cytb* nucleotide sequences. The blood samples were collected from 58 unrelated birds during the marking in the nests. We have performed a comparative analysis of newly obtained data on microsatellites and *Cytb* with existing data. Genetic differentiation analysis between different native populations of French Pyrenees, Croatia and Israel identified two genetic clusters that differentiate populations from the Balkan and Iberian Peninsulas. Genetic diversity analysis based on microsatellites demonstrated similar levels among all populations while analysis of *Cytb* detected somewhat lower diversity in the population from Serbia. Further analyses demonstrated that all analyzed populations experienced a recent bottleneck event. Phylogeographic analysis based on *Cytb* sequences showed that the most frequent haplotype is found in all Griffon vulture populations and that each population possesses private haplotypes. Considering the serious recent bottleneck event which the

*Corresponding author, e-mail: slobodan.davidovic@ibiss.bg.ac.rs

population from Serbia experienced we estimated the overall relatedness among the birds from this population. The level of inbreeding was relatively high and on average it was 8,3% while the mean number of relatives for each bird was close to three. Our data suggest that, even though a relatively high level of inbreeding can be detected among the individual birds, the Griffon vulture population from Serbia can be used as a source population for restocking and reintroduction programs in the region. The observed genetic differentiation between the populations from the Iberian and Balkan Peninsula suggest that the introduction of foreign birds should be avoided and that local birds should be used instead.

Keywords:
population genetics, microsatellites, *Cytb* sequencing, genetic diversity, phylogeography, inbreeding

Multivariate chemometric modeling coupled with Raman spectroscopy for paprika varieties discrimination

Stefan M. Kolašinac, Ilinka Pećinar, Zora P. Dajić Stevanović

Faculty of Agriculture, University of Belgrade, Nemanjina 6, 11000 Belgrade, Serbia

The main goal of this paper is to find the most appropriate multivariate classification model which can classify paprika varieties in physiological stage of maturation (deep red stage) analyzed by Raman microspectroscopy. Several multivariate chemometric classification models, such as PCA-LDA (Principal Component Analysis - Linear Discriminat Analysis), PCA-QDA (Principal Component Analysis - Quadratic Discriminat Analysis), PLSDA (Partial Least Square Discriminat Analysis) and SVM (Support Vector Machines) are performed to determine the model which best fits with target variety of paprika. Before making prediction models pre-processing part is done, including baseline correction, normalization and the second derivative calculation of spectra. According Raman spectra, the several characteristic bands were obtained, including those at 1511-1519, 1149-1157 and 998-1006 cm^{-1}, all assigned to carotenoids with 9 conjugated double bonds in main polyene chain (e.g. capsorubin and capsanthin). The data are divided into the training (3/4 of samples) and validation (1/4 of samples) data. All tested classification chemometric models showed high rate of discrimination accuracy (between 80-100 %) in training and validation data, but SVM had the best prediction power. Obtained results might be explained by use of different algorithm compared with the other tested models.

Keywords:
multivariate prediction models, big data, raman spectroscopy, carotenoids

*Corresponding author, e-mail: stefan.kolasinac@agrif.bg.ac.rs

Gene expression pattern in Edward syndrome: A bioinformatic analysis on what creates significant low life expectancy

Supantha Dey[1,2]

[1] *Department of Genetic Engineering and Biotechnology, University of Dhaka, Nilkhet Road, Dhaka-1000, Bangladesh.*
[2] *Pine Biotech, New Orleans, LA, USA.*

Abstract

Edwards syndrome or trisomy 18 is a form of autosomal aneuploidy. It occurs due to an additional copy of chromosome 18 (complete or partial), which results in severe birth defects and, ultimately, early infant death. But other forms of autosomal aneuploidy like the Down Syndrome have a relatively longer lifespan. Even with a prevalence rate of 1 in 3000 children, there are not many studies on the genetic profile of this disease. Also, available treatment guidelines are not able to increase a patient's life expectancy significantly. This study used public datasets from GEO which used oligonucleotide microarrays (Affymetrix, U133 Plus 2.0) and analyzed whole genome expression profiles in amniocytes (AC) and chorion villus cells (CV) from pregnancies with normal karyotypes and aneuploidies. Bioinformatic analysis of trisomy 18 patients' sample was completed by comparing it with the control group's sample. It helped us to identify the expression patterns of the genes associated with the increased prevalence of Edwards syndrome. This study elucidates how the abnormal expression of these genes can affect vital bodily functions and are correlated with major factors that increase the prevalence of early infant deaths due to respiratory abnormalities, central apnea, cardiovascular diseases, etc.

At first, we selected the significantly expressed genes by applying differential expression in edgeR. Threshold was set for p value at <0.05, FDR <0.001, logFC >1.5. Also, a gene ontological study was done. It revealed how some crucial pathways associated with cellular metabolism were affected. Most significantly, carbohydrate and fat metabolism rates were altered. Genes related to DNA transcription and/or repair were severely downregulated. Mainly FGFR2, BMP6, GABRA 2, NF-IB, HACD4, ZNF597, etc. genes were altered, and these genes are associated with autoimmune diseases, bone marrow necrosis, renal failure, encephalopathy, and mental retardation. Our findings indicated that these altered genes are mainly responsible for reduced life quality and expectancy. Our study suggested these altered genes might be used in genome editing technology and genetic engineering in future for developing treatment protocols to reduce early mortality rate as literature study has also supported. Overall, this study could improve the overall life quality of trisomy-18 patients.

Keywords:

Trisomy 18, biomarkers, gene expression regulation and alternation, bioinformatics, infant mortality.

*Corresponding author, e-mail: shupanthodey@gmail.com

Overall proteome variability manifested by alternative splicing in the dorsal striatum samples of mouse chronic social conflict model

Vladimir Babenko*, Dmitry Smagin, Irina Kovalenko, Anna Galyamina, and Natalia Kudryavtseva

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

Abstract

Mouse model of social conflict emerged in 1991 features repeated fighting experience affecting gene networks performance along with behavioral abnormalities. Within model framework we maintained three groups (6 species per group): a) the controls; 2) mice with 20 days experience of aggression; 3) group 2) after period of fighting deprivation (14 days).

Using RNA-Seq routine we outlined the genes in dorsal striatum that manifested the largest protein-coding variability mediated by alternative splicing (AS) events identified by rMATs software. The maximum protein coding transcripts per gene (from 12 down to 4) sample was featured by 188 genes. We applied Gene Ontology (GO) enrichment annotation routine (string-db.org) to elucidate the major functions enrichment in this gene pool. We obtained highly interconnected network with number of edges 233 vs expected average 126 (FDR<1E-16). There were 59 genes connected with nervous system development (FDR<1.8E-12), including neural development and generation of neurons GO categories. Neuron projection category enrichment outlined 49 genes (FDR<1.2E-14).

Interestingly, both aggressive and defeated groups were featured by downturn of neuron development and axonogenesis processes compared to control and deprived samples.

Another major GO category of AS related enrichment was synaptic genes set maintaining their plasticity upon stress response. Notably, by comparing aggressive vs. defeated groups in synaptic plasticity, we found that defeated group maintain preferably short isoforms (exon skipped) upon stress, while, oppositely, the long isoforms in aggressive groups prevail over skipped ones for synaptic genes. At the same time, approximately 60% of AS events were identical between aggressive and defeated groups.

We also confirmed that the majority (more than 70% in defeated samples and 60% in aggressive samples) of alternative coding exons maintain frame 0 (from 3 possible: 0,1,2) codons ratio, while non-coding poison exon frames don't differ from that of constitutively spliced ones (about 50% of frame 0). It implies distinct mode of evolution of poison vs coding alternative exons: exon decay vs exon shuffling as an assumption.

Keywords:

chronic social conflict model, RNA-Seq, alternative splicing stress response, dorsal striatum, brain stress response.

Acknowledgment:

The study was funded by Russian Science Foundation 19-15-00026.

*Corresponding author, e-mail: bob@bionet.nsc.ru

Analyzing biological network using grah theory

Xhilda Merkaj[1], Eglantina Kalluçi[2], Darjon Dhamo[3]

[1,2] Department of Applied Mathematics, Faculty of Natural Sciences, University of Tirana
[3]Automation Department, Faculty of Electrical Engineering, Polytechnic University of Tirana
Mother Teresa Square, Tirana, Albania, 1001

Abstract
Networks can represent a great variety of different types of data. The nodes can represent different entities (such are proteins or genes in biological networks) and edges provide information about the links between the nodes. Our focus in this article is on graph theory methods for biological networks. Firstly, we are going to introduce the basic graph theory concepts and the various graph types. In addition, we will survey methods and approaches in graph theory, along with current applications in biological networks. Our intention is to analyze graph theory concepts, methods and models to real biological networks. Furthermore, we intend to develop an application on Matlab to do all our analysis, so even someone who has no information about Matlab can perfectly use the application and perform all the analysis he/she is interested in. This application will cover all the network properties and just with one click everyone will have the results ready. We expect this application to reach a very broad spectrum of users varying from experts to beginners, while encouraging them to enhance it further.

Keywords:
graph theory, graph models, biological networks, Matlab applications.

*Corresponding author, e-mail: xhilda.merkaj@fshn.edu.al

Inhibitor potential of some Au(III) complexes against SARS-CoV-2 Spike Glycoprotein: A Molecular Docking Study

Žiko Milanović[1*], Ana Kesić[1], Dejan Milenković[1], Edina Avdović[1] and Zoran Marković[1]

[1] *University of Kragujevac, Institute for Information Technologies, Jovana Cvijića bb, 34000 Kragujevac, Serbia*

Abstract

The pharmacologic properties of gold compounds have been known since the end of the 19th century. In the last decade, gold complexes have received increased attention due to the variety of their applications. Square planar Au(III) complexes are suitable candidates for biological investigation because of useful substances with a good stability profile. Some Au(III) complex could be significantly stabilized, even at neutral pH, with the appropriate choice of the inert ligands, preserving its peculiar biological properties. In this paper, the molecular interactions between active binding sites of SARS-CoV-2 Spike Glycoprotein and analyzed compounds, Au(III) complexes ([Au(terpy)Cl]2+ (C1) and [Au(bipy)Cl2]+ (C2),), hydroxychloroquine and chloroquine, were investigated by molecular docking simulations. The crystal structure of investigated receptor (PDB ID: 6VSB) was extracted from RCSB Protein Data Bank in PDB format. The binding affinity of investigated compounds was examined by the AutoDock 4.2 software. AutoDockTools was used to calculate the Kollman partial charges and addition polar hydrogens. The flexibility of the ligands was considered, while the protein kept on as the rigid structure in the ADT. The Lamarckian Genetic Algorithm (LGA) method was used for protein-ligand flexible docking. The parameters for the LGA method were determined as follows: a maximum number of energy evaluations is 250,000, a maximum number of generations is 27,000, and mutation and crossover rates are 0.02 and 0.8, respectively. The pockets and binding sites of the target receptor were determined by the AutoGridFR program. The binding energies of the docked compounds of C1 and C2 against SARS-CoV-2 Spike Glycoprotein were found to be in the range between -28.5 and -23.5 kJ/mol, as opposed to hydroxychloroquine and chloroquine which are -34.5 and -35.8 kJ/mol, respectively. The obtained results revealed that C1 and C2 bind at the same binding pockets to SARS-CoV-2 Spike Glycoprotein, as well as hydroxychloroquine and chloroquine, by weak non-covalent interactions. The most prominent interactions are hydrogen bonds, alkyl-π, and π-π interactions. The preliminary results suggest that Au(III) complexes showed good binding affinity against SARS-CoV-2 Spike Glycoprotein, as evident from the free binding energy (ΔGbind in kJ/mol) and that might exhibit inhibitory activity against SARS-CoV-2 Spike Glycoprotein.

Keywords:

Au(III) complex, terpyridine, bipyridine, molecular docking, SARS-CoV-2 Spike Glycoprotein

Acknowledgments:

Authors acknowledge the Ministry of Education, Science and Technological Development of the Republic of Serbia for the support through Contracts No. 451-03-9/2021-14/200378.

*Corresponding author, e-mail: ziko.milanovic@uni.kg.ac.rs

MAIN ORGANIZER

Vinča Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade

MAIN CO-ORGANIZERS

Faculty of Biology, University of Belgrade

Faculty of Mathematics, University of Belgrade

Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

Mathematical Institute of SASA, Belgrade

Serbian Society for Bioinformatics and Computational Biology

CO-ORGANIZERS

BioIRC – Bioengineering Research and Development Center

Faculty of Agriculture, University of Belgrade

Faculty of Engineering, University of Kragujevac

Faculty of Medicine, University of Belgrade

Faculty of Sciences and Mathematics, University of Niš

Faculty of Sciences, University of Novi Sad

Institute for Biological Research "Siniša Stanković", Belgrade

Institute for General and Physical Chemistry, Belgrade

Maize Research Institute Zemun Polje

Serbian Genetic Society

Serbian Society for Molecular Biology

BELGRADE BIOINFORMATICS CONFERENCE 2021 HAS BEEN SUPPORTED BY

Ministry of Education and Science of the Republic of Serbia

International Centre for Genetic Engineering and Biotechnology

Factory World Wide

Seven Bridges Genomics

INDEX

CIP – Каталогизација у публикацији
Библиотека Матице српске, Нови Сад 57
BIOLOGIA Serbica / editor-in chief Milica
Matavulj. –
Vol. 34, no. ½ (dec. 2012) - . – Novi Sad : Faculty of
Sciences, Department of Biology and Ecology,
2012-. –
Ilustr. ; 28 cm
Dva puta godišnje. – Je nastavak: Зборник
радова Природно-математичког факултета.
Серија за биологију = ISSN 0352 – 1788
ISSN 2334 – 6590 = Biologia Serbica
COBISS.SR – ID 275725831

ISSN 2334-6590
9 772334 659001

2021

Vol. 43 - No. 1 Special Edition

Biologia Serbica

ISSN 2334-6590