ISSN 2334-6590
UDK 57 (051)

Biologia Serbica

Book of Abstracts
Belgrade BioInformatics Conference 2018

Volume 40 - No. 1
(Special Edition)
2018

BELBI 2018

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

Content

International Advisory Committee

Prof. Vladik Avetisov, The Semenov Institute of Chemical Physics, Russian Academy of Sciences, Moscow, Russia
Prof. Vladimir Brusic, Li Dak Sum Chair Professor of Computer Science, University of Nottingham Ningbo, China
Prof. Radu Constantinescu, Department of Physics, University of Craiova, Romania
Prof. Oxana Galzitskaya, Institute of Protein Research, Russian Academy of Sciences, Moscow, Russia
Prof. Nikolay A. Kolchanov, Institute of Cytology and Genetics, Novosibirsk, Russia
Prof. Sergei Kozyrev, Department of Mathematical Physics, Steklov Mathematical Institute, Moscow, Russia
Prof. Yuri Manin, Max Planck Institute for Mathematics, Bonn, Germany
Prof. Matteo Marsili, Quantitative Life Sciences Department, International Centre for Theoretical Physics, Trieste, Italy
Prof. Zoran Obradovic, Computer and Information Sciences Department, Temple University, Pennsylvania, USA
Dr Zoran Ognjanovic, Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
Prof. Sergey Petoukhov, Mechanical Engineering Research Institute, Russian Academy of Sciences, Moscow, Russia
Prof. Natasa Przulj, Department of Computer Science, University College London, UK
Prof. Paul Sorba, Laboratory of Theoretical Physics and CNRS, Annecy, France
Prof. Bosiljka Tadic, Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, Slovenia
Prof. Vladimir Uversky, Department of Molecular Medicine, University of South Florida, USA
Prof. Sergey Volkov, Bogolyubov Institute for Theoretical Physics, National Academy of Sciences, Kiev, Ukraine

International Program Committee

Dr Bojana Banovic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
Prof. Jan Baumbach, Technical University of Munich, Germany
Dr Milos Beljanski, Institute of General and Physical Chemistry, University of Belgrade, Serbia
Sanja Brdar, Institute BioSense, Novi Sad, Serbia
Prof. Alexandre de Brevern, University Paris Diderot, Sorbonne Paris Cite, France
Prof. Marko Djordjevic, Faculty of Biology, University of Belgrade, Serbia
Prof. Branko Dragovich, Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
Dr Valentina Djordjevic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
Prof. Dragan Matic, Faculty of Sciences, Department of Mathematics and Informatics, University of Banja Luka, Bosnia and Herzegovina
Prof. Nenad Mitic, Faculty of Mathematics, University of Belgrade, Serbia
Dr Ivana Moric, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
Prof. Gordana Pavlovic-Lazetic, Faculty of Mathematics, University of Belgrade, Serbia
Dr Ana Simonovic, Institute for Biological Research " Siniša Stanković", University of Belgrade, Serbia
Prof. Antonio Starcevic, Faculty of Food Technology and Biotechnology, University of Zagreb, Croatia
Prof. Alessandro Treves, International School for Advanced Studies, Trieste, Italy
Dr Nevena Veljkovic, Vinča Institute of Nuclear Sciences, University of Belgrade, Serbia
Prof. Igor V. Volovich, Department of Mathematical Physics, Steklov Mathematical Institute, Russian Academy of Sciences, Moscow, Russia
Prof. Evgeny Zdobnov, Swiss Institute of Bioinformatics, Switzerland

Local Organizing Committee

Dr Bojana Banovic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
Dr Milos Beljanski, Institute of General and Physical Chemistry, University of Belgrade, Serbia
Prof. Branko Dragovich, Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia
Dr Lidija Djokic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr Valentina Djordjevic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
Prof. Marko Djordjevic, Faculty of Biology, University of Belgrade, Serbia
Jelena Guzina, Faculty of Biology, University of Belgrade, Serbia
Dr Nikola Kotur, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr Jovana Kovacevic, Faculty of Mathematics, University of Belgrade, Serbia
Dr Natasa Kovacevic-Grujicic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr Jelena Kusic Tisma, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Prof. Sasa Malkov, Faculty of Mathematics, University of Belgrade, Serbia
Mirjana Maljkovic, Faculty of Mathematics, University of Belgrade, Serbia
Prof. Nenad Mitic, Faculty of Mathematics, University of Belgrade, Serbia
Dr Ivana Moric, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
Prof. Gordana Pavlovic-Lazetic, Faculty of Mathematics, University of Belgrade, Serbia
Prof. Zeljko Popovic, Faculty of Sciences, University of Novi Sad, Serbia
Dr Jelena Samardzic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr Biljana Stankovic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr Branko Tomic, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
Dr Nevena Veljkovic, Vinča Institute of Nuclear Sciences, University of Belgrade, Serbia

PREFACE

On behalf of the Local Organizing Committee, I have a great pleasure to welcome you to Belgrade at the second **Belgrade BioInformatics Conference – BelBi2018.**

Several research institutions, faculties and scientific societies from Serbia have joined their forces to organize this international conference focused on different aspects of bioinformatics. The Conference is organized by the Institute of Molecular Genetics and Genetic Engineering University of Belgrade (the main organizer) and University of Belgrade Faculty of Mathematics, University of Belgrade Faculty of Biology, Mathematical Institute of SASA, and Vinča Institute of Nuclear Sciences University of Belgrade, as co-organizing institutions, in cooperation with several other institutions from Serbia, and two COST (European Cooperation in Science and Technology) Actions, BM1405 and CA15120.

The overarching idea of the BelBi2018 Conference is to illuminate different aspects of bioinformation systems, including theoretical approaches in modeling different phenomena in life sciences, information technologies necessary for analysis and understanding huge amount of generated data, application of computer science and informatics in precision medicine, microbiology, and plant science, search for novel remedies against debilitating diseases and drug development, and other similar topics. Therefore, the Conference aims to encourage the cooperation and exchange of ideas between scientists in related fields, with a special emphasis on regional networking.

This 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 to prof. Zeljko D. Popovic, Managing Editor, for all his effort to bring this enormous work successfully to the finish.

I would like to thank to 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 Abdus Salam International Centre for Theoretical Physics (Trieste, Italy) that supported attendance of regional participants. And at last, but not the least, the Local Organizing Committee is very grateful to all Conference's sponsors - Telekom Srbija, Vivogen d.o.o., Qiagen GmbH, Labena d.o.o., Alfa Genetics d.o.o., Kefo d.o.o., DSP Chromatography d.o.o., ProMedia d.o.o., and RTC d.o.o., with the hope that they will be with us for many years to come.

At the end, I would like to wish you to have a memorable time in our city, one of the oldest cities in Europe that embraces and welcomes everyone

More information about the Conference can be found at the web site http://belbi.bg.ac.rs/.

Belgrade, June 2018
dr Ivana Morić
On behalf of BelBi2018

Conference program

The conference is focused on three main research fields including the following topics.
Other subjects related to the main themes are welcome.

Theoretical Approaches to BioInformation Systems (TABIS)

- Structure and function of DNA, RNA and proteins
- Gene expression and the genetic code
- Neurons and cognition
- Biological networks

Bioinformatics and Data Mining of Biological Data (BDMBD)

- Bioinformatics and data analysis in biomedicine, genetics, microbiology and plant science
- Sequencing technologies and data analysis
- Structure and molecular interaction predictions
- Data analysis and integration for OMICs data
- Data mining methods, algorithms, and applications in life sciences and precision medicine
- Big data and data science
- Data analytics, pattern recognition and machine learning in data analysis
- Bioinformatics applications for OMICs data
- Predictive Models for OMICs data
- Bioinformatics databases and algorithms

Biomedical Informatics (BI)

- Translational Bioinformatics
- Disease Models & Epidemiology
- Predictive Modeling and Analytics in Healthcare
- Biomedical Imaging and Data Visualization
- Biomedical/Health database integration and management
- Biomedical data/text mining

SUNDAY, JUNE 17th

Location: Lobby of Hotel "Palace" (Toplicin venac 23)

17.00–20.00 Early Registration

MONDAY, JUNE 18th

9.00–14.00 Registration & Opening ceremony

Location: Lobby of Rectorate of the University of Belgrade (Studentski trg 1)

9.00–12.00 Registration

Location: Congress Hall of Rectorate of the University of Belgrade (Studentski trg 1)

Chairs:

Prof. Branko Dragovich, Mathematical Institute of the Serbian Academy of Sciences and Arts

Prof. Gordana Pavlović-Lažetić, Faculty of Mathematics, University of Belgrade

Prof. Nenad Mitić, Faculty of Mathematics, University of Belgrade

Prof. Marko Đorđević, Faculty of Biology, University of Belgrade

Dr Ivana Morić, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

10.00–10.30 Welcome Speeches

10.30–11.15 **prof. AlessandroTreves**, International School for Advanced Studies, Italy

Cortical mechanisms underlying language processes

11.15–12.15 Welcome Cocktail (*Atrium of the University of Belgrade Rectorate*)

12.15–13.00 **prof. Alexandre Morozov,** Rutgers, the State University of New Jersey, USA

Biophysical Models of Chromatin Plasticity and Nucleosome Dynamics

13.00–13.45 **prof. Predrag Radivojac**, Indiana University, USA

Beyond Patterns: Deep Understanding of Biology with Machine Learning

Location: COST CA15120 Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

9.00–14.00 COST CA15120 WG Meeting, OpenMultiMed

14.00–15.00 Lunch*

(Hotel "Palace", Toplicin venac 23)

*Lunch will be served every day during the Conference from 13.30 until 15.30

15.00–18.00 Afternoon Sessions

Location: TABIS Session, Conference Hall, Hotel "Palace" (Toplicin venac 23)

15.00–15.25 **Marsili** – Relevance

15.25–15.50 **de Brevern** – Analysis of allosteric effect of pathologic variants at the light of local protein conformations

15.50–16.15 **Tadic** – Why Human Brain Networks are Hyperbolic

Location: COST CA15120 Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

15.00–16.15 COST CA15120 WG Meeting, OpenMultiMed

MONDAY, JUNE 18th

16.15–16.45 Afternoon Coffee Break

Location: TABIS Session, Conference Hall, Hotel "Palace" (Toplicin venac 23)

16.45–17.10 **Fimmel** – On Dichotomy Classes and Bijections of the Genetic Code

17.10–17.35 **Struengmann** – Circular Codes in the Genetic Information

17.35–18.00 **Kozyrev** – Biology is a constructive physics

Location: COST CA15120 Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

16.45–17.00 Introduction by the STSM Committee

17.00–17.20 **Demirkol** – Characterization of novel risk predictors in colorectal cancer

17.20–17.40 **Lazareva** – Network-constrained bi-clustering of patients and multi-scale omics data

17.40–18.00 **Vilor** – Independent Multifactorial Association algorithm to assess genetic and neuroimaging features associated with neurodevelopmental domains

TUESDAY, JUNE 19th

Location: Hotel "Palace" (Toplicin venac 23)

8.00–9.00 Registration

9.00–11.00 Plenary Lectures

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

9.00–9.30 **Severinov** – Maintenance of plasmid DNA at conditions of countering CRISPR-Cas interference

9.30–10.00 **Baumbach** – Network-based disease classification and de novo endophenotyping

10.00–10.30 **Stres** – The rare layers of significance

10.30–11.00 Morning Coffee Break

11.00–14.00 Morning Sessions

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

11.00–11.25 **Vilar** – Inference and prediction in molecular biological systems

11.25–11.50 **Ciliberto** – Cells proliferating under constant checkpoint activation: adaptation, refractory state and memory Opening ceremony

11.50–12.15 **Studholme** – Comparative genomics of a recently emerging epidemic on banana in Sub-Saharan Africa.

12.15–12.35 **Vlahovicek** – Predicting Disease from Gut Microbiota Codon Usage Profiles

12.35–13.00 **Pojskic** – Digital reconstruction of partial DNA profiles of human skeletal remains

13.00–14.00 **Poster Session 1**

1.P1 **Saveljic** – 3 DSimulation of Inflammatory Process in Coronary Arteries

1.P2 **Nikolic** – Geometry Optimization of Nitinol Stent Design based on FEA Topology Optimisation

1.P3 **Vulovic** – Numerical Simulation of Blood Flow and Plaque Progression in Right Femoral Artery Bypass Patient-Specific Case

1.P4 **Nikolic** – Modelling of Monocytes Behaviour inside the Bioreactor

1.P5 **Djorovic** – Parametric Modelling and Computational Examination of Bicuspid Aortic Valve

1.P6 **Grbic** – Conditional Random Fields based approach for classification of the reactants in some metabolic reactions

1.P7 **Stankovic** – Binding of metal ions and water molecules to nucleic bases

1.P8 **Milanovic** – Inactivation of free radical species with selected triazoles

1.P9 **Avdović** – Molecular docking study on the interaction of human procalcitonin with 3-(1-(2-mercaptoethylamino)ethylidene)-chroman-2,4-dion

1.P10 **Doric** – PalFin: A Software Tool to Identify Specific Palindrome Motifs in mtDNA

TUESDAY, JUNE 19th

Location: COST CA15120 Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

11.00–14.00 COST CA15120 MC Meeting, OpenMultiMed

14.00–15.00 Lunch (Hotel "Palace")

15.00–18.30 Afternoon Session

Location: BiDMBD Session, Conference Hall, Hotel "Palace"

15.00–15.15 **Djordjevic – A Kolmogorov**-Smirnov based approach for predicting bacterial transcription targets

15.15–15.30 **Guzina** – Predicting CRISPR/Cas associated small RNAs and their role in bacterial virulence

15.30–15.40 **Chervontceva** - The role of mRNA secondary structure in the control of translation and mRNA degradation in E. coli

15.40–15.50 **Krause** – Establishing Benchmark Criteria for Single Chromosome Bacterial Genome Assembly

15.50–16.00 **Rodic** – Investigating the role of key features in CRISPR-Cas system regulation

16.00–16.15 **Discussion**

Location: COST CA15120 Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

15.00–15.15 **Lautizi** – Extracting survival-relevant subnetworks from multi-scale omics data with KeyPathwayMiner

15.15–15.35 **Brdar** – Clustering and classification of human microbiome data: evaluating impact of different settings in bioinformatics workflows

15.35–15.55 **Belik** – Modelling Hospital Infection Spread in the Polish Regional Healthcare Network

15.55–16.15 **Casas Guijarro** – 3D reconstruction of cerebral reactive oxygen species formation as a gold standard for future in vivo molecular imaging approaches

16.15–16.45 Afternoon Coffee Break

Location: Sponsor Session 1, Conference Hall, Hotel "Palace"

16.45–17.15 Sarah Nema (QIAGEN) QIAGEN Bioinformatics – The journey so far

Location: BiDMBD Session, Conference Hall, Hotel "Palace"

17.15–17.40 **Savic-Pavicevic** – Identifying modifiers of somatic instability and age at onset in myotonic dystrophy type 1 by modeling genetic data

17.35–17.45 **Vukovic** – Higher-order genetic interactions in prostate cancer and benign prostatic hyperplasia

17.45–17.55 **Nikolic** - GARLIC: A Bioinformatic Toolkit for Etiologically Connecting Diseases and Cell Type-Specific Regulatory Maps

17.55–18.05 **List** – Genome-wide endogenous RNA networks highlight novel biomarkers in cancer

18.05–18.15 **Kosvyra** - Developing an Integrated Genomic Profile for Cancer Patients Utilizing NGS Data

18.15–18.25 **Marjanovic** - Bioinformatics pipeline used for Next Generation Sequencing analysis of predictive markers in hematological malignancies

18.25–18.35 Discussion

Location: COST CA15120 Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

16.45–17.15 COST CA15120 MC Meeting, OpenMultiMed

Location: BiDMBD & BI Sessions, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

17.15–17.40 **Exarchos** – New concepts for the stratification of patients with carotid artery disease: Multiscale modelling and big data analytics

17.40–17.50 **Radovic** – Application of Machine Learning Algorithms to Detect Coronary Artery Disease using Genomic Data

17.50–18.00 **Isailovic** – Numerical simulation of stent deployment procedure in patient specific coronary artery

18.00–18.10 **Simic** – Application of multiscale smeared finite element model for modelling of mass transport in capillary systems and biological tissue

18.10–18.20 **Andjelkovic Cirkovic** – Prediction of Pharmacological Treatment for Patients with Coronary Artery Disease

18.20–18.30 Discussion

WEDNESDAY, JUNE 20th

Location: Hotel "Palace" (Toplicin venac 23)

8.00–9.00 Registration

9.00–11.00 Plenary Lectures – COST BM1405 Session

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

9.00–9.10 Introduction

9.10–9.40 **Rost** – Implications of dark proteome for precision medicine

9.40–10.05 **Tossato** – Computational resources for the study of intrinsically disordered proteins

10.05–10.30 **Ventura** – Combining structural aggregation propensity and stability predictions to re-design protein solubility

10.30–11.00 Morning Coffee Break

11.00–14.00 Morning Sessions

Location: COST BM1405 Session, Conference Hall, Hotel "Palace" (Toplicin venac 23)

11.00–11.30 **Bateman** – Pfam: 20 years of classifying protein repeat families

11.30–11.55 **Tompa** – Phase separation in ALS demonstrates the role of RNA binding and low-complexity regions in collective protein functions

11.55–12.20 **Galzitskaya** – Influence of homo-repeats on the aggregation properties of proteins from 122 proteomes and codon usage in DNA

12.20–12.45 **Mitic** – Disorder predictors precision and accuracy - a computer science view

12.45–13.05 **Discussion**

13.05–14.00 Poster Session 2

2.P1 **Popovic** – Video-based extraction of movement artifacts in electrogastrography signal

2.P2 **Milicevic** – Muscle model with net of fibers used for modelling cell migration

2.P3 **Vulovic** – Finite Element Analysis of the Modified Hip Implant Surface

2.P4 **Ferouka** – Discrete simulation of electrospinning jet's evolution

2.P5 **Marovac** – Classification of proteins into COG categories based on n-gram patterns

2.P6 **Zivanovic** – Optimization of Electrochemical Parameters for Detection of microRNA: Computer Simulation and Experimental Study

2.P7 **Sustersic** – Numerical simulation of electrospinning using PAK and ANSYS software

2.P8 **Jovanovic** – Subtle transcriptomic signals in circulation after myocardial infarction might indicate the ventricular remodeling outcome

2.P9 **Djukic** – Parallelization of software for stent deployment inside artery

2.P10 **Anic** – Neural Networks Implemented on Aorta with Abdominal Aneurism

Location: BiDMBD Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

11.40–12.05 **Grzybowski** – Genetic portrait of Central- and Eastern European populations

12.05–12.30 **Kovacevic-Grujicic** – Insights into the mitochondrial gene pool of Serbian population: phylogenetic and phylogeographic analysis

12.30–12.40 **Kacprowski** – Circulating miRNAs as Potential Liver-Related Biomarkers

12.40–12.50 **Babenko R.** – FTO haplotyping underlines high obesity risk for European populations

12.50–13.00 Discussion

14.00–15.00 Lunch (Hotel "Palace")

15.00–17.15 Afternoon Session

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

15.00–16.30 **ROUND TABLE:** Experience, Challenges and Networking

WEDNESDAY, JUNE 20th

16.30–16.45 Afternoon Coffee Break

Location: Sponsor Session 2, Conference Hall, Hotel "Palace" (Toplicin venac 23)

16.45–17.15 **Alexander Kirpiy** (Vivogen) – Ion Torrent™ bioinformatics environment

GALA DINNER (to be announced)

THURSDAY, JUNE 21st

Location: Hotel "Palace" (Toplicin venac 23)

8.00–9.00 Registration

9.00–11.00 Plenary Lectures

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

9.00–9.30 **Gelfand** – From computer to a test tube. How comparative genomics informs molecular biology

9.30–10.00 **Milosavljevic** – ClinGen Allele and Evidence Registries catalyze the emergence of an open ecosystem of data and knowledge about genetic variants in humans

10.00–10.30 **Brusic** – Classification of Blood Cell Subtypes using Single Cell Gene Expression Data

10.30–11.00 Morning Coffee Break, Conference Hall, & Meet the expert Sponsor Session (Alexander Kirpiy, Vivogen), TV Hall

11.00–14.00 Morning Sessions

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

11.00–11.25 **Przulj** – Mining the Integrated Connectedness of Biomedical Systems

11.25–11.50 **Mamitsuka** – Data-Integrative Machine Learning for Bioinformatics

11.50–12.15 **Hong-Yu OU** – Identification of type II toxin-antitoxin loci in bacterial genome

12.15–12.35 **Stanojevi**c – Use of phylogenetics in the study of viral genomes

12.35–13.05 **Starcevic** – Semantic Ion Vectors – deep learning applied to mass spectrometer

13.05–14.00 Poster Session 3

3.P1 **Pavlovic** – Positional Biases of the Experimentally Characterized T-cell Epitopes

3.P2 **Trifunovic** – Determination of hTERT promoter methylation status using methylation specific PCR

3.P3 **Gemovic** – Function Annotation Algorithm Based on Sequence Spectral Features: Evaluation on Human Transcription Factors

3.P4 **Sumonja** – Exploring the usefulness of graph properties in protein protein interaction predictions

3.P5 **Zelic** – Development Of Anatomically Correct Human Mandible Finite Element Model From CT-Scans

3.P6 **Zukic** – FINDbase: worldwide database for clinically relevant genomic variation allele frequencies

3.P7 **Dragicevic** – Review of (effective) data collection methods for data-driven personalized medicine

3.P8 **Kotur** - Population pharmacogenomic aspect of glucocorticoids response in Serbian population

3.P9 **Mitic Potkrajac** – Risk prediction of bladder cancer progression from gene expression data

3.P10 **Davidovic** – Phylogenetic analysis of putative Balkan specific mtDNA lineages

Location: Location: BiDMBD Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

11.40–12.05 **Banovic-Djeri** – Bioinformatics in plant genetics from molecular biologists point of view

12.05–12.30 **Ivanova** – Sequencing and annotation of resurrection plant Haberlea rhodopensis cp and mt genomes

12.30–12.40 **Paunovic** – ragp: An R toolbox for mining plant Hydroxyproline rich glycoproteins

12.40–12.50 **Yadav** – Characterization of cellular metabolic response in abiotic stress-induced growth in Arabidopsis thaliana utilizing data mining

12.50–13.00 Discussion

THURSDAY, JUNE 21st

14.00–15.00 Lunch (Hotel "Palace")

15.00 – Free Time / Excursion (to be announced)

FRIDAY, JUNE 22nd

9.00–11.00 Plenary Lectures

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

9.00–9.30 **Dragovich** – p-Adic Genetic Code

9.30–10.00 **Petoukhov** – The New Wide Class of Symmetries in Long DNA-Texts. Elements of Quantum-Information Genetics

10.00–10.30 **Volkov** – The Mechanisms of DNA Genetic Information Deactivation in Ion Cancer Therapy

10.30–11.00 Morning Coffee Break

11.00–14.00 Morning Sessions

Location: Conference Hall, Hotel "Palace" (Toplicin venac 23)

11.00–11.25 **Martins** – Spectral signature of gene family trees

11.25–11.50 **Orlov** – Analysis of Differential Alternative Splicing and Gene Networks by RNA-seq Data in Brain Areas of Laboratory Rats

11.50–12.15 **Hofestadt** – Pnlib-Shell: Modeling and Simulation of biological networks based on Petri nets

12.15–12.40 **Zukic** – Bioinformatics strategy for rare disease diagnostics in the era of next-generation sequencing

13.05–14.00 Poster Session 4

4.P1 **Blagojevic** – Defining dynamical property observables which ensure efficient restriction-modification systems establishment in bacterial host

4.P2 **Tasic** – Application of bioinformatics in the identification of autohtonous bacterial strains of Vranjska Banja thermal springs based on different methods

4.P3 **Malkov** – Correlation of intrinsically disordered protein regions content with environmental characteristic in Archaea and

4.P4 **Loncar-Turukalo** – Improving Clustering Performance in Microbiome Studies

4.P5 **Katic** – Motor Imagery Classification using H2O Machine Learning Platform

4.P6 **Jelovic** – RepeatPlus – program for finding repeats in nucleic acids and proteins

4.P7 **Menon** – In silico identification of Transcription End Sites in Human Genome

4.P8 **Maljkovic** – Analysis of Amino Acid Interactions Based on Geometric Distances

4.P9 **Bjelica** – Unobtrusive Human Activity Recognition

4.P10 **Vrecl** – Combined in silico and experimental approach to identify the peptide mimetic of the nanobody that stabilize functional conformational state of the beta2 adrenergic receptor (β2AR)

14.00–15.00 Lunch (Hotel "Palace")

15.00–18.30 Afternoon Session

Location: BI Session, Conference Hall, Hotel "Palace" (Toplicin venac 23)

15.00–15.10 **Perovic** – Prediction of Human Phenotype Ontology Terms for Intrinsically Disordered Proteins

15.10–15.20 **Davidovic R** – DiNGO: stand-alone application for GO and HPO term enrichment

15.20–15.30 **Vinterhalter** – Bioinformatics analysis of correlation between protein function and intrinsic disorder

15.30–15.40 **Oros** – MALDI-TOF/TOF and diagnosis of bacterial UTIs

15.40–15.50 **Milosevic** – Information extraction from tables: case studies on extracting demographic information from tables in clinical trial literature and drug-drug interactions from tables in drug labels

15.50–16.00 **Milovanovic** – Numerical approach for determination of virtual functional assessment index in coronary artheries

16.00–16.15 Discussion

FRIDAY, JUNE 22nd

Location: TABIS Session, Banquet Hall, Hotel "Palace" (Toplicin venac 23)

15.00–15.25	**Babenko** – Brain regions transriptome analysis in mouse chronic stress model
15.25–15.35	**Dushanov** – Effect of mutant NMDA receptors on oscillations in a model of hippocampus
15.35–15.45	**Graovac** – Investigating interplay of intracellular regulation and population dynamics in a bacterial restriction-modification system
15.45–15.55	**Misic** – Arithmetical Regularities Inside the Standard Genetic Code as a Clue for the Investigation of Natural Biocomputing
15.55–16.05	**Păuna** – Reduction method for reaction-diffusion equations from biology
16.05–16.15	Discussion
16.15–16.45	Afternoon Coffee Break
16.45–17.15	**CLOSING CEREMONY** (Conference Hall, Hotel "Palace")
17.15–18.00	Meeting of the Serbian Society for Bioinformatics and Computational Biology (Conference Hall, Hotel "Palace")

INVITED SPEAKERS

Intrinsic metric of sparse random ensembles

Vladik Avetisov

N. N. Semenov Institute of Chemical Physics of the Russian Academy of Sciences, Kosygina 4, 119991 Moscow, Russia

Abstract:
Biologists often deal with multidimensional, combinatorially large spaces of states. When considering natural selection from the viewpoint of molecular biology, we introduce a sequence space with an enormous number of sequences. Indeed, even for a short genome only 1 Kb long, we have a 103-dimensional sequence space with such a large number of possible sequences that we use abstract values, such as 101000, for illustration. However, it is known that any physically realizable ensemble represents only a vanishingly small subset of states of such large spaces. Given this situation, the question then arises whether a rare random subset of a multidimensional space is representative of that space. In particular, is the metric of a rare random subset of a Euclidean space always a Euclidean metric?

Intuitively, it might seem that this statement is always answered in the affirmative. I wish to demonstrate, however, that this situation is not always the case. I will consider three counterexamples. One of these counterexamples is a geometrical example wherein a branching random process generates a sparse subset of points in a multidimensional Euclidean space. Another example is topological and relates to network systems. For a random network, a sparse subset is given by the rare links between the network nodes, and the spectral density of random sparse adjacency matrices is investigated. The third example relates to the so-called "fractal" or "crumpled" polymeric globule that is a very popular modern model of tight packing of long DNA chains in a cell nucleus. Via this example, a map of rare contacts inside a polymeric globule is considered.

These three examples show that, independent of the nature of the system, sparse ensembles possess a characteristic ultrametric structure of hierarchically nested classes of equivalence specified with respect to the distances between the ensemble elements.

Keywords:
combinatorially large spaces, sparse ensembles, ultrametric, hierarchy.

This work has been partially funded within frameworks of the state task for ICP RAS 0082-2014-0001 (state registration #AAAA-A17-117040610310-6).

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

Brain regions transriptome analysis in mouse chronic stress model

Vladimir Babenko*, Roman Babenko, and Natalia Kudryavtseva

Institute of Cytology and Genetics SB RAS, Lavrentieva 10, 630090 Novosibirsk, Russia

Abstract:
Animal model of stress is a highly preferred method for elucidating the mechanisms of stress response to environmental factors. Recent studies in plant, and animal stress models proved that expression of small and long non-coding RNA (miRNA and lncRNA), along with the posttranscriptional mRNA regulation, such as alternative splicing, are the major response factors in living systems. Aside from response factors, it is of ultimate importance to elucidate the major molecular processes and gene networks which alter their expression/splicing profiles upon the stress invocation.

Using brain regions RNA-Seq data for a murine model of chronic stress for 5 brain regions (Hypothalamus, Hippocapus, Striatum, Ventral Tegmental Area, Midbrain Raphe Nuclei) and 3 social behavior groups (control, 'winner', 'loser') , we approached the task of assigning the brain regions the most highly expressed and gene region discriminative genes. We found that the major genes networks incorporating such genes are: a) metabolic b) neuroendocrine response; c) neurogenesis; d) synaptogenic. The first two networks are featured by hypothalamus gene region, while the neurogenesis genes network is characteristic for hippocampus, and striatum features the Darpp32 gene involved in dopamine mediated gene network response. Alongside, the Differentially Alternatively Spliced (DAS) genes were identified and the corresponding networks were elucidated.

We manifest that each brain region features a specific genes set underlined by chronic stress maintenance. A big portion of DAS genes was featured by homeostatic balance sustain connected with metabolic and signal networks dynamics. Also, the corresponding genes networks were confirmed based on RNA-Seq data co-variation analysis. The specifics and biases of the brain specific RNA-Seq analysis features will be discussed.

Keywords:
bioinformatics, data mining, computer science, nucleic acids, sequencing, brain region, chronic stress animal model

This work has been funded by RSF14-15-00063.

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

Bioinformatics in plant genetics from molecular biologists point of view

Bojana Banović Đeri*[1], Dragana Dudić[2], and Vesna Pajić[2]

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

[2]*Faculty of Agriculture, University of Belgrade, Nemanjina 6, Zemun, Serbia*

Abstract:
Nowadays with the development and broad use of NGS technology in genetics research we tend to produce a large amount of data, much greater than we can process and/or store. This in return causes a heightened need for improved utilization of the existing data. Data science, which emerged as a new discipline over the last decade, proved to be a highly useful in revealing hidden information and unknown relations between the existing data, like associations between genotype and phenotype variants in molecular genetics. In plant sciences it presents a valuable asset providing guidelines for fighting against plant pathogens/abiotic stresses, stabilizing and/or increasing plants' yield, suggesting new breeding designs, etc. However, the problem of continuous production of new biological data and new pipelines for their analysis, without extracting all available valuable information from the data that we already have by using methods that already exist, continues to persist. We will discuss the synergy between molecular biology and bioinformatics using the case studies of model and non-model plants and how two disciplines should guide and support each other in order to prevent continuous data/pipelines hyper production, while improving the usage of the existing data/tools to extract the hidden knowledge.

Keywords:
NGS technology, data science, plant genetics, hidden knowledge

This work has been funded by MESTD, Republic of Serbia (project 173005).

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

Pfam: 20 years of classifying protein repeat families

Alex Bateman

European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, Cambridgeshire, CB10 1SD, UK

Abstract:
The Pfam protein families database was founded over 20 years ago. In the intervening years it has been adding protein families at the rate of about 1,000 per year. Since its foundation Pfam has been classifying protein repeat families. In this talk I will discuss the challenges associated with repeat classification and the often pragmatic decisions that we have taken. The evolution of the HMMER software has had a major impact on our ability to classify repeats and I will describe how this has changed over time. I will finish with prospects for creating a complete and accurate classification of repeat families

Keywords:
bioinformatics, data mining, computer science, nucleic acids, sequencing

This work has been funded by EMBL-EBI.

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

Network-based disease classification and de novo endophenotyping

Jan Baumbach

Technical University of Munich, Maximus-von-Imhof-Forum 3, 80333 München, Germany

Abstract:
On major obstacle in current medicine and drug development is inherent in the way we define and approach diseases. Here, we will discuss the diagnostic and prognostic value of (multi-)omics panels in general. We will have a closer look at breast cancer subtyping and treatment outcome, as case example, using gene expression panels - and we will discuss the current "best practice" in the light of critical statistical considerations. Afterwards, we will introduce computational approaches for network-based medicine. We will discuss novel developments in graph-based machine learning using examples ranging from Huntington's disease mechanisms via lung cancer drug target discovery back to where we started, i.e. breast cancer subtyping and treatment optimization - but now from a systems medicine point of view. We conclude that systems medicine and modern artificial intelligence open new avenues to shape future medicine.

Keywords:
drug development, multi-omics, future medicine

This work has been funded by H2020, DFG and VILLUM Foundation.

*Corresponding author, e-mail: jan.baumbach@wzw.tum.de

Classification of blood cells by applying pattern recognition techniques to single cell gene expression data

Vladimir Brusic

Faculty of Science and Engineering, University of Nottingham, 199 Taikang East Road, University Park, Ningbo 315100, China

Abstract:
Advances in biotechnology, instrumentation and miniaturisation have enabled the insights into genetic material of single cells. Single cell transcriptomics (SCT) enables simultaneous measurements of mRNA concentrations of thousands of genes from tens of thousands of cells. The ability to profile transcriptomes of individual cells increases the diagnostic potential of transcriptomic studies as compared to the earlier studies of bulk samples. Unfortunately, the sensitivity of mRNA measurement from single cells is markedly lower than from bulk samples to the extent that only a fraction of actual transcripts that do exist in a single cell will be captured and identified.

Peripheral blood mononuclear cells (PBMC) contain lymphocytes (T cells, B cells, and NK cells), monocytes, and a very low number of dendritic cells. Each of these cell types have many subtypes defined by the expression of different combination of cellular markers. These subtypes correlate to multiple cellular dimensions including differentiation stage, tissue location, activation history and disease states. Sophisticated cell sorting techniques, such as flow cytometry are used for separation of target subsets of cells, with the purity routinely exceeding 99%. Flow cytometry is used in a range of medical applications (e.g. clinical trials, pathology, immunology, haematology, transplantation), and for research including transcriptomic analysis of sorted samples.

The limitation of SCT is a limited sensitivity of transcript capture. The CD8+ T cells (cytotoxic lymphocytes) express CD8 surface marker that is a dimer consisting of CD8A and CD8B chains. CD8A is captured expressed in only 35% of CD8+ T cells, CD8B in 74%, and CD8A or CD8B in 75% of CD8+ T cells. CD4 is a major surface marker of T helper cells, but CD4 transcript is found by SCT only in 5-10% of cells sorted by CD4-specific antibodies. Furthermore, CD8A transcript is found in 5% of NK cells. Standard surface markers, therefore, cannot be used for accurate classification of cell subtypes from SCT experiments. We have identified more than 1,000 transcriptomic candidate markers of cell types and subtypes from PBMCs. Furthermore, we determined patterns that classify major cell types within PBMCs and their subtypes with high accuracy. These patterns were assessed against cell marker knowledge and show consistency with the known markers of cell types. Testing with several independent data sets confirmed accuracy of classifications. Our results indicate that pattern recognition based analysis of a single SCT run can produce cell clustering results that would otherwise require tens of cell sorting experiments. The ability to accurately identify cell type of individual cells from PBMCs is essential for improvement of disease marker discovery.

Keywords:
single cell classification, pattern recognition, transcriptomics signatures

This work has been funded by EMBL-EBI.

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

Cells proliferating under constant checkpoint activation: adaptation, refractory state and memory

Andrea Corno[1], Claudio Vernieri[1], and Andrea Ciliberto[1*]

[1] *IFOM (The Firc Institute of Molecular Oncology) Via Adamello 16, 20139, Milan, Italy*

Abstract:
The Spindle Assembly Checkpoint (SAC) arrests cells in metaphase, until conditions for proper chromosome segregation are met. Cells under a constant activation of the SAC are arrested in metaphase, until they eventually adapt (we call them, adapting cells). Adaptation occurs in the presence of an active checkpoint and has the properties of a stochastic process. After adaptation, cells either die or resume proliferation, regardless impaired chromosome segregation. Here, we have examined the cell cycle of cells after adaptation (we call them, adapted cells). We show that when the SAC is induced by overexpressing one of its components, adapted cells divide faster than adapting cells (i.e., they become refractory to the SAC). The refractory state persists in the transient absence of the SAC (up to 3 hours). Adapted cells are larger than normal cycling cells, although our data suggest that the refractory state cannot be explained with the achievement of a critical size. Mass spectrometry analysis showed that adapted cells differ in many respects compared to cycling cells, from metabolism to cell cycle control. In agreement with this observation, we identified synthetic lethalities typical of adapted cells.

Keywords:
quantitative biology, adaptation, cell division

This work has been funded by AIRC, the Italian Association for Cancer Research.

*Corresponding author, e-mail: andrea.ciliberto@ifom.eu

Analysis of allosteric effect of pathologic variants at the light of local protein conformations

Alexandre G. de Brevern[1,2,3,*]

[1] *Biologie Intégrée du Globule Rouge UMR_S1134, Inserm, Univ. Paris Diderot, Sorbonne Paris Cité, Univ. de la Réunion, Univ. des Antilles, F-75739 Paris, France*
[2] *Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.*
[3] *Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.*

Abstract:
Proteins are highly dynamic macromolecules. To analyze their inherent flexibility, computational biologists often use molecular dynamics (MD) simulations. The quantification of protein flexibility is based on various methods such as Root Mean Square Fluctuations (RMSF) that rely on multiple MD snapshots or Normal Mode Analysis (NMA) that rely on a single structure and focus on quantifying large movements.

Alternative in silico approaches assess protein motions through the protein residue network or dynamical correlations from MD simulations. An alternative yet powerful approach based on small prototypes or "structural alphabets" (SAs) can be used. SAs approximate conformations of protein backbones and code the local structures of proteins as one-dimensional sequences. Protein Blocks (PBs) are one of these SAs.

Applying PB-based approaches to biological systems such as the DARC protein, the human αIIb β3 integrin and the KISSR1 protein highlighted the major interest of PBs in understanding local deformations of large protein structures. Specifically, these analyses have shown that a region considered as highly flexible through RMSF quantifications can be seen using PBs as locally highly rigid. This unexpected behavior is explained by a local rigidity surrounded by deformable regions. To go further, we used PBs to analyze long-range allosteric interactions in the Calf-1 domain of αIIb integrin and will also present new unpublished results. Our tool named PBxplore allows the analyses of MDs in terms of PBs and quantification with entropy index and also different types of visualization.

The presentation of this research will be done in 5 successive parts: (i) the presentation and interest of the PBs, (ii) the MD of the integrins, (iii) the analysis of the results using PBs with the PBxplore software, (iv) recent developments based on PBxplore to compare reference and variant proteins, i.e. providing information about potential allosteric events and (v) the extension of the analyses to the disordered protein.

Keywords:
bioinformatics, computer science, protein structures, entropy, biostatistics, disorder proteins, long-range interactions

This work has been funded by grants from the Ministry of Research (France), University Paris Diderot, Sorbonne, Paris Cité (France), National Institute for Blood Transfusion (INTS, France), National Institute for Health and Medical Research (INSERM, France) and labex GR-Ex. The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program "Investissements d'avenir" of the French National Research Agency, reference ANR-11-IDEX-0005-02. AdB acknowledges Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grant (number 5302-2).

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

p-Adic Genetic Code and Bioinformation

Branko Dragovich[1,2]

[1] *Institute of Physics, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia*
[2] *Mathematical Institute, Serbian Academy of Sciences and Arts, Kneza Mihaila 36, 11000 Belgrade, Serbia*

Abstract:
From mathematical point of view, the genetic code (GC) is a map from the set of 64 codons onto the set of 20 amino acids and one stop signal. The standard GC is almost unique in all living organisms with respect to a huge number of mathematical possibilities. The vertebrate mitochondrial (VM) code is relatively simple and other dozens of genetic codes can be considered as its slight variations. In the VM code, an amino acid is coded by one, two or three codon doublets. When a doublet codes the same amino acid, one can say that codons in the doublet are close (similar) in the informational sense. In the last ten years was shown that the p-adic distance is a simple and adequate mathematical tool to describe codon closeness (similarity).

p-Adic distance (p is a prime number) is the most employed example of ultrametrics and plays a central role in p-adic analysis and its applications in modeling physical and biological systems with hierarchical structure. In p-adic approach to modeling genetic code it is enough to use p-adic distance between some integers. In fact, p-adic distance is related to divisibility of difference between two integers with respect to prime number p: larger divisibility – smaller distance. p-Adic distance between any two integer numbers is smaller or equal to 1. When p-adic distance is smaller, the related numbers are p-adic closer.

In p-adic modeling of the VM code we assign 64 natural numbers in the form to the 64 codons appropriately identifying nucleotides in codons with digits in these numbers. In particular, we take the following identification: C (Cytosine) = 1, A (Adenine) = 2, U (Uracil) = T (Thymine) = 3, and G (Guanine) = 4. With respect to the smallest 5-adic distance between codons one obtains 16 quadruplets. Each of these quadruplets splits into two doublets under 2-adic distance. Then each of these 32 doublets, which contain two p-adically closest codons, codes one amino acid or stop signal.

This p-adic approach for codons is extended to amino acids. From p-adic point of view the genetic code is an ultrametric network. This approach is also used to investigate possible evolution of the genetic code. p-Adic distance is a useful tool for study similarity between biomolecular sequences, which elements are nucleotides, codons or amino acids.

Keywords:
p-adic genetic code, bioinformation, similarity between sequences, ultrametric network

This work has been funded by MESTD, Republic of Serbia (projects 173052 and 174012).

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

New concepts for the stratification of patients with carotid artery disease: Multiscale modelling and big data analytics

Themis Exarchos[1,2*], Vasiliki Potsika[1], Michalis Mantzaris[1], Dimitrios I. Fotiadis[1]

[1] Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, University Campus of Ioannina, GR45110, Ioannina, Greece
[2] Dept. of Informatics, Ionian University, Plateia Tsirigoti 7, GR49100, Corfu, Greece

Abstract:
Carotid artery disease, the primary trigger of ischaemic cerebrovascular events including stroke, causes major morbidity, mortality and healthcare costs worldwide. Still, treatment is based on criteria established in the 90s that do not take into account the molecular evolution we have witnessed since, nor the introduction of new medication, leading to remarkably high unnecessary surgical treatment while missing most patients at risk.

TAXINOMISIS will provide novel disease mechanism-based stratification for carotid artery disease patients to address the needs for stratified and personalised therapeutic interventions in the current era. This will be achieved through (1) the dissection of mechanisms mediating carotid artery disease, and identification of susceptibility and protection factors of plaque erosion and/or rupture using longitudinal cohorts and multi-omics, (2) the definition of distinct disease phenotypes and endotypes, and generation of molecular fingerprints of high versus low-risk states through systems medicine, (3) the development of a multilevel risk prediction model of the symptomatic plaque incorporating new biomarkers and advanced imaging, implemented in a software, to assist patient stratification and clinical decision making, (4) the development of novel pharmacogenomics solutions based on lab-on-a-chip technology to support personalized treatment, (5) the evaluation of the new risk prediction model and lab-on-a-chip device in a prospective observational clinical study, and (6) the assessment of regulatory, cost-effectiveness and ethical issues towards the implementation and commercialization of the programme's outcomes.

TAXINOMISIS has therefore the potential to rationally change the current state-of-the-art in the stratification of patients with carotid artery disease by reducing unnecessary operations, refining medical treatment and opening up new avenues for therapeutic intervention, while strengthening the European biotechnology sector.

Keywords:
patient stratification, carotid artery disease, multilevel prediction models

This work has been funded by Horizon 2020 TAXINOMISIS project, GA 755320.

*Corresponding author, e-mail: exarchos@cc.uoi.gr

On Dichotomy Classes and Bijections of the Genetic Code

Elena Fimmel

Faculty of Computer Sciences, Mannheim University of Applied Sciences, Paul-Wittsack-Str. 10, 68163 Mannheim, Germany

Abstract:
In this talk a recently developed concept of a binary dichotomic algorithm (in short BDA) for analyzing the genetic code will be presented. The genetic code is based on four nucleotide bases that are combined into triples called codons. BDAs decompose the set of 64 possible codons into two equally sized subsets, called dichotomic classes. The concept arose as a generalization of the classical dichotomies, e.g. the biologically important Rumer´s dichotomy. It will be shown that when the codon/anticodon complementarity is understood as a dichotomy, it can be computed in the same algorithmic way the Rumer's classes and the parity dichotomic classes are obtained in the recent non-power model of the genetic code. In addition, the global framework of bijective transformations of the nucleotide bases is discussed and we clarify when dichotomic partitions can be generated. Dichotomic classes allow the uncovering of many symmetry properties of the genetic code and mirror operations that have a factual biological function in the decoding center of the ribosome.

Keywords:
genetic code, Rumer's classes, dichotomy, bijective transformations

*Corresponding author, e-mail: e.fimmel@hs-mannheim.de

Influence of homo-repeats on the aggregation properties of proteins from 122 proteomes and codon usage in DNA

Oxana V. Galzitskaya[1*], Georgii S. Novikov[2], Nikita V. Dovidchenko[1], Mikhail Yu Lobanov[1]

[1] *Institute of Protein Research, Institutskaya Street 4, Pushchino, Russia*
[2] *St. Petersburg Academic University, Khlopina str., 8/3, St. Petersburg, Russia*

Abstract:
The influence of homo-repeats with lengths larger than four on the aggregation properties of proteins has been studied across 122 eukaryotic and bacterial proteomes. It has been found that proteins with homo-repeats are on average longer than in the whole database. The ability of proteins with homo-repeats to aggregate cannot be explained only by the presence of long homo-repeats in the proteins. There should be other characteristics of proteins increasing the aggregation property including such as appearance of homo-repeats in pairs in the same protein.

We have found the biases for codon usages for some amino acids in homo-repeats for 122 proteomes and for all amino acids when the same codon is used for homo-repeats. Therefore, the codon usage biases are not organism or taxa specific. Length-dependent codon usage is observed only for lysine homo-repeats. Similar results are obtained for human proteins with homo-repeats associated with diseases. Moreover, for proteins associated with diseases, the fraction of proteins for which the same codon is used for homo-repeats is larger than for proteins from the 122 proteomes. We are the first who demonstrated for human proteome that in some cases the splicing sites correspond to the homo-repeats and these sites more often appear at the C-terminal part of the homo-repeats.

Keywords:
homo-repeat, codon usage, proteome, splicing site, disease

This work has been funded by the programs "Molecular and Cellular Biology" (01201353567) and by the Russian Science Foundation.

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

From computer to a test tube. How comparative genomics informs molecular biology

Mikhail S. Gelfand[1,2,3,4]

[1] *Kharkevich Institute for Information Transmission Problems RAS, Bol'shoi Karetnyi per. 19, 127051 Moscow, Russia*
[2] *Skolkovo Institute of Science and Technology, Center for Data-Intensive Biomedicine and Biotechnology, 1 Nobel Street, 143028 Moscow, Russia*
[3] *Higher School of Economics, Faculty of Computer Science, 3 Kochnovsky Proezd, 125319 Moscow, Russia*
[4] *Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Leninskiye Gory GSP-1 MSU 1-73, 119991 Moscow, Russia*

Abstract:
Comparative analysis of genomic and transcriptomic data allows one to formulate non-trivial hypotheses that greatly simplify experimental work: instead of looking for a needle in a haystack, without even knowing whether there is any needle, or whether it is a needle and not a pin, or an entire sewing kit, or a manufacturing robot, the biologists validates a specific conjecture. What is important, such hypotheses may go far beyond mere similarity-based functional annotations, and often suggest novel biological phenomena. I will present several recent examples of such studies.
The first study started with the analysis of gene composition of sugar-catabolism loci. One such locus, recently described as Escherichia coli sulphoquinovose degradation system yihTSV, was observed to be similar to the lactose catabolism locus of Bacilli. It lead to the suggestion that this locus may have a dual function, encoding at the same time the second lactose utilization system, complementing the classical Jacob-Monod lactose operon. In addition, promoters and binding sites for CRP and adjacent DeoR-family transcriptional regulator YihW were determined. Experiments demonstrated that one of the three promoter for yihT is active only during growth on lactose, and that transcription of yihT is controlled by YihW hat acts as a local regulator opposite to the global regulator of sugar metabolism, CRP.
Two more similar studies stemmed from a systematic analysis of transcriptional regulators of the GntR family, their regulons and binding motifs.
Two closely related regulators, UxuR and ExuR, had been believed to regulate catabolism of hexuronates binding to overlapping sets of DNA sites. However, the data from a large-scale ChIP-Seq suggested that ExuR might be a global regulator. Indeed, it has been shown to influence expression of a number of other transcription factors, thus serving as the master regulator of catabolism. Further comparative analysis is needed to characterize the evolution of this regulatory network.
One of the loci predicted to be controlled by UxuR contained the gene yjjM, encoding a transcription factor. This prediction was validated in an experiment. Subsequent ChIP-seq analysis of YjjM revealed a modulon enriched in genes involving in mobility and formation of biofilm. Indeed, direct experiments demonstrated that ΔyjjM mutants formed large colonies (were more motile) and were less capable of forming biofilms. This link is intriguing, as the hexuronate metabolism has been linked to colonization of the intestine.
Such approaches are not limited to prokaryotes. Polypedilum vanderplanki is a chironomid that can survive almost complete desiccation. The seral transcriptome during the dehydration-rehydration cycle was compared with the transcriptome of a congeneric desiccation-sensitive midge Polypedilum nubifer, and a set of specific, differentially expressed genes was identified. Computational analysis revealed a motif TCTAGAA enriched in upstream regions of these genes. It closely resembles the binding motif of drosophila heat shock transcription activator (Hsf). In vitro cell-line experiments mimicking the dehydration-rehydration cycle proved that Hsf knockdown suppressed activation of the desiccation-related genes and lead to a five-fold reduction in cell survival. In the hindsight, the link between desiccation and heat shock is almost obvious, as the physiological effects such as protein misfolding and DNA damage are similar. However, this is a striking example of an evolutionary expansion of a pre-existing regulatory system to a new set of genes in response to an emerging, unique stimulus.

Keywords:
comparative genomics, transcriptomics, ChIP-Seq, carbohydrate metabolism, lactose, hexuronates, regulation of transcription, desiccation, heat shock

This work has been funded by RSF grants 14-50-00150 and 18-14-00358.

*Corresponding author, e-mail: gelfand@iitp.ru

IGenetic portrait of Central- and Eastern European populations

Tomasz Grzybowski[1*], Boris A. Malyarchuk[2], Natasa Kovacevic-Grujicic[3], Milena Stevanovic[3], The NEXT Consortium

[1] Nicolaus Copernicus University, Collegium Medicum in Bydgoszcz, Department of Forensic Medicine, Curie-Skłodowskiej Str. 9, Bydgoszcz, Poland
[2] Institute of Biological Problems of the North, Genetics Laboratory, Portovaya Str. 18, Magadan, Russia
[3] University of Belgrade, Institute of Molecular Genetics and Genetic Engineering, 444a Vojvode Stepe, Belgrade, Serbia

Abstract:
Recent developments in massively parallel sequencing (MPS) technologies have certainly sped up human DNA analyses undertaken for population genetic and phylogeographic purposes. Nevertheless, high resolution genetic data from populations of Central and Eastern Europe are still relatively scarce. In this presentation, all available results from our research into both haploid and autosomal DNA markers from Slavic-speaking populations will be discussed on the basis of comparable data available from other populations of Western and Eastern Eurasian ancestry. Special emphasis will be placed on the origin and molecular dating of mitochondrial subhaplogroups of potential Central- and Eastern European regional specificity. Initial steps towards identifying autosomal ancestry informative markers capable of discriminating between Central European populations will also be outlined. These results will be interpreted in light of the data on the formation of a Slavic identity emerging from other genetic studies, as well as other disciplines, including archaeology, linguistics, history and physical anthropology.

Keywords:
exome sequencing, mitogenome variability, ancestry informative markers, Slavic-speaking populations

This work has been funded by The National Centre for Research and Development (Poland).

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

PNlib-Shell: Modeling and simulation of biological networks based on Petri nets

Ralf Hofestädt

University Bielefeld, Technical Facult, Universitätsstr. 25, 33615 Bielefeld, Germany

Abstract:
Based on a short introduction of Petri nets and an overview of existing simulation shells this presentation will focus to a new Petri net simulation shell based on the OpenModelica software tool. A user interface will be presented, which allows the access to the Petri net library (PNlib) of OpenModelica. The PNlib-Shell provides a powerful simulation environment for the simulation of biological networks. Based on this new shell Petri net models can be easily created, simulated and analyzed. In addition the new system includes basic features to check and evaluate the model and to analyze simulation results generated by the simulation back-end.

Keywords:
bioinformatics, modeling and simulation, biological networks, Petri net

This work has been funded by the Volkswagen Stiftung (A115842).

*Corresponding author, e-mail: ralf.hofestaedt@uni-bielefeld.de

Sequencing and annotation of resurrection plant *Haberlea rhodopensis* cp and mt genomes

Zdravka Ivanova[2], Evelina Daskalova[1], Gergana Zahmanova[1,3], Elena Apostolova[1,2], Galina Yahubyan[1,2], Ivan Minkov [2,3] and Vesselin Baev[1,2*]

[1] *Department of Plant Physiology and Molecular Biology, University of Plovdiv, 24 Tsar Assen St.,4000 Plovdiv, Bulgaria*
[2] *Institute of Molecular Biology and Biotechnology, 105 Ruski Boulevard, 4000 Plovdiv, Bulgaria*
[3] *Center of Plant Systems Biology and Biotechnology, 139, Ruski Boulevard, 4000 Plovdiv, Bulgaria*

Abstract:
Drought affects more people than any other natural disaster. It has been estimated that half of the world population will live in areas of high water scarcity by 2020 that will likely pose an increasingly dramatic problem of food shortage.

Haberlea rhodopensis is a paleolithic tertiary relict species best known as resurrection plant with remarkable tolerance to desiccation. *Haberlea rhodopensis* exposed to severe drought stress shows the ability to maintain the structural integrity of the photosynthetic apparatus which re-activaties easily upon rehydration. In addition to its homoiochlorophyllous nature, *H. rhodopensis* with such capability of resurrection is a trait of significant importance in the global climate change.

The NGS sequencing and annotation analysis of *H. rhodopensis* cp genome uncover several intriguing features which can be used as a base to understand the resurrection tolerance of this plant. Specifically, the *H. rhodopensis* cp genome harbors 137 genes, of which 86 protein-coding. The site-specific selection analysis point out positively selected sites in several chloroplast genes such as atpE, rbcL, psbI, psbA, ndhH and accD. The observed specific cp genomic features of *Haberlea rhodopensis* may be interpreted as being a consequence of molecular adaptation to drought stress, which awards an evolutionary advantage to this species.

We are currently working on the annotation of the mt genome of *Haberlea rhodopensis*, so far, we were able to identify 61 genes including 3 rRNA, 23 tRNA, and 35 protein-coding genes. The organelle genomes reported in this study will allow to understand the function of the specific sites under selection by developing site-directed mutagenesis assays or by developing point mutations at those sites, which can be a step and bridge to bioengineering of drought-tolerant crops.

Keywords:
sequencing, genome assembly, resurrection plant, Haberlea rhodopensis

This work has been funded by the programs "Molecular and Cellular Biology" (01201353567) and by the Russian Science Foundation.

*Corresponding author, e-mail: vebaev@plantgene.eu

Insights into the mitochondrial gene pool of Serbian population: phylogenetic and phylogeographic analysis

Nataša Kovačević-Grujičić[1*], Slobodan Davidović[1], Boris Malyarchuk[2], Jelena Aleksić[1], Tomasz Grzybowski[3], Miroslava Derenko[2], Andrey Litvinov[2], and Milena Stevanović[1,4,5]

[1] *Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11010 Belgrade, Serbia*
[2] *Institute of Biological Problems of the North, Russian Academy of Sciences, 18 Portovaya Street, 685000 Magadan, Russia*
[3] *Ludwik Rydygier Collegium Medicum, Faculty of Medicine, Nicolaus Copernicus University, Marii-Sklodowskiej-Curie Str. 9, 85-094 Bydgoszcz, Poland*
[4] *Faculty of Biology, University of Belgrade, Studentski Trg 16, 11000 Belgrade, Serbia*
[5] *Serbian Academy of Sciences and Arts, Knez Mihailova 35, 11000 Belgrade, Serbia*

Abstract:
The Balkan Peninsula, which connects Asia Minor and Europe, played an important role during the entire history of humans because it served as an important migration corridor for the initial peopling of Europe during the early Upper Paleolithic as well as for post-glacial recolonization of Europe and for numerous subsequent human migrations. Since Serbian population inhabits the central part of the Balkan Peninsula, this population may harbour genetic traces of past complex demographic processes in the Balkans.

We have demonstrated previously that mtDNA (sub)haplogroup composition and frequency distribution in Serbians fits typical European maternal landscape. Our next goal was to comprehensively study complete mitogenomes of Serbians, a population linking westward and eastward South Slavs, in order to gain a better insight not only into the time and place of the origin of particular mtDNA lineages but also into the events that might have had a role in shaping the Serbian maternal gene pool. For that purpose, we completely sequenced 164 mitogenomes detected in Serbian population belonging to mtDNA (sub)haplogroups that have different frequency distributions in European populations (H, HV, V, U, K, W, I, X, N1a, N1b, R0a, L2a1, and D4). Subsequently, they were analysed phylogenetically against available complete mtDNAs of modern and ancient Eurasians.

Phylogenetic and phylogeographic analyses showed that Serbians share a number of mtDNA lineages with Southern, Eastern-Central and North-Western Europeans, harbouring also lineages found in populations from Middle East/Caucasus and East Asia. MtDNA subhaplogroups of probable Southern European origin, as well as putative Balkan-specific lineages, were delineated. In particular, Serbian population shares mtDNA lineages with West and East Slavs as well as with Germanic populations from Northern and Central Europe.

We found that the exceptional diversity of contemporary maternal gene pool of Serbians may be related to genetic inputs of two main sources: 1) indigenous pre-Slavic Balkan populations whose maternal gene pool was affected by migrations of various populations over time (e.g. Yamnaya) and 2) Slavic and Germanic newcomers in the early Middle Ages.

Keywords:
mitochondrial DNA, phylogeny, Balkan Peninsula, Serbian population

This work has been funded by MESTD, Republic of Serbia (project 47025), Russian Foundation for Basic Research grant number (16-34-00014) and Faculty of Medicine, CM UMK, Poland, (grant number MN-4/WL/2016)

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

Biology is a constructive physics

Sergei Kozyrev

Steklov Mathematical Institute, Gubkina 8, 119991 Moscow, Russia

Abstract:
Yuri Manin's approach to Zipf's law (Kolmogorov complexity as energy) is applied to investigation of biological evolution. Model of constructive statistical mechanics where complexity is a contribution to energy is proposed to model genomics. Scaling laws in genomics are discussed in relation to Zipf's law.

Keywords:
bioinformatics, computer science, Kolmogorov complexity

This work has been funded by RNF 14-50-00005.

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

Data-Integrative Machine Learning for Bioinformatics

Hiroshi Mamitsuka[1, 2]

[1] *Bioinformatics Center, Institute for Chemical Research, Kyoto University*
Gokasho, Uji 6110011, Japan
[2] *Department of Computer Science, School of Science, Aalto University, Konemiehentie 2, Espoo 02150, Finland*

Abstract:
Modern biology has generated a large number of data, covering wide aspects in biology, while this diversity of biological data reveals a possibility of combing totally different data for the same objective. For example, gene function can be inferred from various information, such as sequence similarities, gene coexpression, connectivity of protein-protein interactions and/or metabolic networks, etc. This would be reasonable in many senses. First different data are usually obtained from different perspectives, which can complement each other. Second, experimentally obtained biological data are noisy, by which missing information of one dataset can be filled by another. The key point of this diverse data is in their various data formats. For example, gene expression is a typical table with rows of genes and columns of experimental conditions, such as time-series, different individuals, etc. This means one gene is a vector of real values. On the other hand, sequences of genes and proteins are strings. Chemical compounds can be molecular graphs. Interactions of protein-protein, drug-target etc. are a network or graph, in which one molecule is a node. Thus the question is how we can integrate such different types of data effectively and efficiently. A naïve, baseline approach would be to transform one data type into another, such that a set of vectors can be transformed into a graph/network by similarity or correlation between two vectors. In this talk, I would like to show machine learning-based approaches for integrating different data formats, which would be more reasonable than the baseline idea, with one or more examples of integrating real biological data. The focus of the talk is unsupervised learning, particularly semi-supervised clustering and collaborative matrix factorization.

Keywords:
bioinformatics, data mining, machine learning, semi-supervised clustering, matrix factorization

This work has been funded by MEXT, Japan.

*Corresponding author, e-mail: mami@kuicr.kyoto-u.ac.jp

Relevance

Matteo Marsili

Abdus Salam ICTP, Strada Costiera 11, Trieste, Italy

Abstract:
Most living organisms needs to cope with an environment that is not under their control and to extract meaningful cues from its virtually infinite dimensional stochastic dynamics, in order to survive and reproduce. Likewise, the brain of higher organisms constructs abstract representations from the stream of raw sensory inputs, on which it builds cognitive maps of objects, places, events and their relation. Similarly, deep neural networks extract relevant features from high dimensional data such as pictures, that enables classification and generalization to remarkable degrees of accuracy.

The problem, in these tasks, is that of achieving optimal representations at a fixed sample size, that in most cases is far from being enough to infer a model.

What are the properties of efficient representations and what characterizes relevant variables or features?

We show that there is a unique way to quantify the information content of a given representation, that we call "relevance". This is the entropy of the frequency with which different observations occur in the sample. Most informative samples are those which maximize the relevance (i.e. the information content of the most efficient representation) and are characterized by a power law distribution of frequencies, where the exponent gauges the tradeoff between resolution and relevance. Zipf's law characterizes samples for which resolution and relevance are traded one for the other at par. In deep neural networks, layers characterized by Zipf's law are those that generalize best. Finally, we show how relevance can be used to identify important residues in proteins and neurons in multi-electrode array data.

Keywords:
bioinformatics, dimensional reduction, information theory

*Corresponding author, e-mail: marsili@ictp.it

ClinGen Allele and Evidence Registries catalyze the emergence of an open ecosystem of data and knowledge about genetic variants in humans

Aleksandar Milosavljevic

Baylor College of Medicine, One Baylor Plaza, 77030 Houston, Texas, USA

Abstract:
The Clinical Genome Resource (ClinGen) is an US NIH-funded project that aims to develop a knowledge base that will inform interpretation of human genetic variation in clinical contexts, including medical genetics and research data that helps decide if a variant observed in a patient is pathogenic. The project recognizes that it is increasingly difficult to aggregate all the information about human genetic variation required for research and clinical applications in a single centralized database or a data warehouse. This problem calls for an easily accessible distributed infrastructure and web-interfaces for aggregation and exchange of variant information that comes from many sources and is accessible to all participants globally. To secure interoperability of components within such distributed infrastructure, a semantically interoperable model for providing the representation of components is necessary. To seed the emergence of an open linked data ecosystem that will help globally integrate and harness variant information, we developed the ClinGen Allele Registry (CAR) and Evidence Registry (ER). These resources are accessible both via a user-friendly web user interface and programmatically via well-documented APIs that facilitate interoperability with existing tools. The CAR provides readily available globally unique variant identifiers that enable aggregation of information from different sources about any variant. A core element of the CAR is a canonicalization service that groups variant identifiers denoting the same nucleotide or amino acid variant and assigns a unique and dereferenceable canonical identifier ("CAid"). The canonicalization service is implemented using a highly optimized in-memory sequence alignment-based index that hosts over 500 thousand known reference sequences (major genome assemblies and transcripts from key resources). More than 650 million distinct variants are currently registered, including those from gnomAD, ExAC, dbSNP, MyVariant.info, COSMIC, and ClinVar, and a smaller number of variants registered by CAR users. The Evidence Registry (ER) complements the CAR by providing means for sharing evidence and interpretations about variant's pathogenicity using ACMG/AMP or similar guidelines. The services are accessible using both via Web UIs and APIs. The ER currently hosts over 5,000 ACMG guideline based variant interpretations imported from multiple sources. For interoperability with curation interfaces and other components that may import to or export from the ER via the ER APIs, the interpretations are represented using the newly developed ClinGen SEPIO semantically interoperable variant interpretation model. We demonstrate the utility of CAR and ER, by (1) showing how the variant information linked by the CAR provides raw material for reasoning about a variant's pathogenicity, (2) demonstrating API-based integration of the ER through the ClinGen Pathogenicity Calculator, where ACMG/AMP-style variant interpretations are imported to / and exported from to ER by a single click; and (3) analyzing over 5,000 ACMG guideline based variant interpretations currently present in the ER.

Keywords:
ClinGen, ClinGen Allele Registry, Evidence Registry, CAid

This work has been funded by the US National Institutes of Health, National Human Genome Research Institute contract U41.

*Corresponding author, e-mail: amilosav@bcm.edu

Disorder predictors precision and accuracy - a computer science view

Nenad Mitić

Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia

Abstract:
It is well known that intrinsically disordered proteins play significant role in various cell processes. Also, it has been proven that a protein's disorder content is related to its function. As the process of experimentally determining disordered regions in proteins is slow and expensive, only a few hundred among millions of known proteins have experimentally confirmed disordered regions. As a consequence, large number of computer programs - disorder predictors - has been developed for prediction of disordered regions in proteins.

As the reasons and mechanism of appearing disordered regions in proteins are not theoretically formally described, very different bases are used for prediction algorithms in disorder predictors. The question that constantly arises is which one of them is better and more accurate. A computational analysis of most commonly used disorder predictors was done based on proteins from DisProt (database of experimentally discovered protein disorder regions) as a benchmark. The analysis includes a comparison of predicted and experimentally found disordered regions, as well as different predictors aspects including precision and accuracy at the points on the timeline associated with the detection of the regions and the construction of the predictors.

Keywords:
disorder predictors, DisProt, accuracy of prediction, proteins

This work has been funded by MESTD, Republic of Serbia (projects 174021 and 44006).

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

Biophysical Models of Chromatin Plasticity and Nucleosome Dynamics

Alexandre V. Morozov

Department of Physics & Astronomy and Center for Quantitative Biology, Rutgers University, 136 Frelinghuysen Road, Piscataway NJ 08854, USA

Abstract:
Genomic DNA in eukaryotes is organized into arrays of nucleosomes. Each nucleosome consists of up to 147 base pairs of DNA wrapped around a histone octamer core. The resulting complex of DNA with histones and other proteins forms a multi-scale, dynamically regulated structure called chromatin. At the most fundamental level of chromatin organization, arrays of nucleosomes form so-called 10-nm fibers which resemble beads on a string and which subsequently fold into higher-order structures. Depending on the organism and cell type, 75-90% of genomic DNA is packaged into nucleosomes. Far from being static entities whose function is to make most of genomic DNA inaccessible to other factors, nucleosomal arrays constantly undergo complex dynamic rearrangements. For example, neighboring nucleosomes may invade each other's territories through DNA unwrapping and translocation, or through initial assembly in partially wrapped states. I will discuss evidence for the surprising plasticity of nucleosomal arrays observed using genome-wide nucleosome maps in baker's yeast S. *cerevisiae* and fruit fly D. *melanogaster*. In particular, I will argue that previously identified nucleosome-depleted regions are in fact covered with unstable nucleosomes in the fly genome. To explain this plasticity and, more generally, predict chromatin structure and its effect on gene regulation, my group has developed a hierarchy of biophysical models based on statistical mechanics of finite-size particles bound to DNA. Our models are in agreement with the observed genome-wide distributions of inter-dyad distances, wrapped DNA lengths, and nucleosome occupancies. Furthermore, they explain *in vitro* measurements of accessibility of nucleosome-covered target sites, and naturally describe nucleosome-induced cooperativity between DNA-binding factors. Overall, our findings indicate that dynamics of nucleosomal arrays plays a key role in multi-scale chromatin organization.

Keywords:
nucleosome positioning, chromatin structure, gene regulation

This work has been funded by the National Institutes of Health (R01 HG004708) and by the Alfred P. Sloan Foundation.

*Corresponding author, e-mail: morozov@physics.rutgers.edu

Analysis of Differential Alternative Splicing and Gene Networks by RNA-seq Data in Brain Areas of Laboratory Rats

Yuriy L. Orlov[1,2*], Anatoly O. Bragin[1], Anton V. Tsukanov[2], Vladimir N. Babenko[1,2]

[1] *Institute of Cytology and Genetics SB RAS, Prospekt Lavrentyeva 10, 630090 Novosibirsk, Russia*
[2] *Novosibirsk State University, 1 Pirogova str, 630090 Novosibirsk, Russia*

Abstract:
This work presents analysis molecular mechanisms of behavior based on omics data in laboratory animal models using computer tools. To analyze genetic component of aggressive behavior RNA-seq data in brain areas of laboratory rats (Rattus norvegicus) were used. Two gray rat lines were selected over more than 30 years at the Institute of Cytology and Genetics SB RAS, Novosibirsk for studying of genetics factors determining aggression behavior. The brain areas were selected due to previous studies indicating their role in inhibitory control of aggression and stress response regulation.

RNA-seq sequencing of rat brain areas samples was performed by Illumina HiSeq 1500. Tophat2 aligner, Trimmomatic tool and Cufflinks program were used for gene expression calculations. The detection of differential gene splicing was performed using rMATS software and Python-based application. The tools for gene coexpression analysis of gene network related to aggressive behavior were used. ANDVisio (Associative Network Design) and ANDSystem tool allow reconstruction of associative gene networks controlling biochemical and physiological processes based on scientific literature mining. The literature on genetics of aggressive behavior suggests multi-loci determination of aggressive behavior without any major gene reported, thus demanding integration of bioinformatics approaches.

Understanding of molecular mechanisms controlling brain function needs analysis of gene splicing events landscape. Grin1 gene as was shown recently is relevant to the stress induced response in neurons. The study reconfirmed the brain specific deviations of gene expression in the RNA-seq data, in particular in synapse specific genes. The systemic research of gene expression in the brain cells using comprehensive experimental approaches is required for interdisciplinary neurobiological studies, and future research has to be conducted using new transcriptome data.

Authors are grateful to K.A. Tabanyukhov, A.L. Markel, I.V. Chadaeva for science discussion.

Keywords:
bioinformatics, transcriptomics, gene networks, behavior, laboratory animals

This work has been funded in part by RFBR grant.

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

Identification of type II toxin-antitoxin loci in bacterial genomes

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 toxin–antitoxin (TA) systems, initially identified with the post-segregational killing function of natural plasmids, are highly abundant on the chromosomes of most free-living bacteria. They are involved in multiple life activities of bacteria, such as abortive infection and persister cell formation. The type II TA system consists of a stable toxin protein and a labile cognate antitoxin protein encoded by a bicistronic locus. Here we report an open-access database of bacterial type II toxin-antitoxin loci, TADB2 (http://bioinfo-mml.sjtu.edu.cn/TADB2/). With the aid of text mining and manual curation, it recorded 6,193 type II TA loci in 870 replicons of bacteria and archaea, including 105 experimentally validated TA loci. In addition, the online tool TAfinder combines the homologue searches and the operon structure detection, allowing the prediction for type II TA pairs.

With TAfinder, 212 putative type II TA loci were identified in ten completely sequenced Klebsiella pneumoniae genomes, including 77 toxin proteins containing Gcn5-related N-acetyltransferase (GNAT) domain. GNAT toxin acetylates aminoacyl-tRNA and blocks protein translation. It is abolished by the cognate antitoxin that contains the ribbon-helix-helix (RHH) domain. We presented the experimental demonstration of a GNAT-RHH TA locus kacTA found in K. pneumoniae HS11286, a strain resistant to multiple antibiotics. Phylogenetic analysis grouped the toxin KacT into a distinct clade with the comparison to the reported Salmonella enterica GNAT toxin. Overexpression of KacT halted cell growth and resulted in persister cell formation. Then, the expression of the bicistronic kacAT locus was up-regulated during antibiotic stress. KacT and KacA formed a heterohexamer that interacted with promoter DNA, resulting in negative autoregulation of kacAT transcription. Finally, the crystal structure of KacAT-DNA complex revealed the forming of a heterohexamer, KacT-KacA2-KacA2-KacT. The direct interaction of KacA and KacT adopts a novel W-shaped architecture with the two KacT molecules at the distant ends. Our structural analysis reveals that the inhibition of the KacT toxic effect is achieved by the break of the functional KacT dimer via four KacA protein binding. In conclusion, TADB 2.0 might provide better support for understanding the important roles of type II TA systems in the prokaryotic life activities.

Keywords:
bacterial toxin–antitoxin system, acetyltransferase-type toxin, Klebsiella pneumoniae, database, prediction tool

This work has been funded by grants from the 973 program, Ministry of Science and Technology, China (2015CB554202) and the National Natural Science Foundation of China (31371261, 31670074).

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

Digital reconstruction of partial DNA profiles of human skeletal remains

Naris Pojskic[1*], Samim Konjicija[2], Dinko Osmankovic[2], Semir Doric[1]

[1] University of Sarajevo - Institute for Genetic Engineering and Biotechnology, Zmaja od Bosne 8 Kampus, 71000 Sarajevo, Bosnia and Herzegovina
[2] University of Sarajevo - Faculty of Electrical Engineering, Zmaja od Bosne 8 Kampus, 71000 Sarajevo, Bosnia and Herzegovina

Abstract:
The success of DNA identification of missing persons or profiling of archaeological skeletal remains largely depends on the ability of an analyst to generate as complete profile as possible. The success of positive DNA match between a missing person and reference samples rapidly decreases with the number of undetected loci regardless of marker type (STRs, SNPs, DNA sequences, etc.). Either small amount of DNA or degraded DNA may be the cause of incomplete/partial DNA profile. Degradation of DNA is the main cause of partial profiling that reduces the chances of positive identification or estimation of haplogroups. Such processes are generally expected in archaeological skeletal remains. The most common problems caused by low number of genome copies are allele drop-out and locus drop-out. The aim of the project is to develop a strategy based on a machine learning approach (neural network and Bayesian network) as well as to apply the maximum likelihood method for estimating missing variants of the DNA locus. The project addresses four type of markers: Y-STR loci, autosomal STR loci, SNPs, and missing DNA fragments. Quality prediction requires quite large dataset, which already exist within a human population genetics database maintained by Institute for Genetic Engineering and Biotechnology, University of Sarajevo. Since the project is in the realisation phase, this presentation includes the results of estimation of missing variants at Y-STR loci. For estimation of missing numbers in a 23-number sequence, an optimal estimation algorithm was used. In order to optimally estimate the values of missing numbers, it was necessary to have a value for comparison and the definition of criteria. Advantage of optimal estimation use is its robustness regarding the „noisy" values in a number sequence. A database of full 23-number sequences was used to determine typical feature value. Based on the determined value, missing numbers are fitted as to minimize the difference between the determined value and the value for the subject sequence. Typical pre-processing and post-processing techniques used in data mining were applied, in order to achieve satisfactory performance of the algorithm regarding the accuracy and error measures.

Keywords:
bioinformatics, machine learning, STR loci, DNA profiles

This work has been funded by Federal Ministry of Education and Science, Government of Federation of Bosnia and Herzegovina (Grant No. 05-39-2498-1/17).

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

Mining the Integrated Connectedness of Biomedical Systems

Natasa Przulj

Computer Science Department, Gower Street, WC1E 6BT London, UK

Abstract:
We are faced with a flood of molecular and clinical data. Various bio-molecules interact in a cell to perform biological function, forming large, complex systems. Large-scale patient-specific omics datasets are increasingly becoming available, providing heterogeneous, but complementary information about cells, tissues and diseases. The challenge is how to mine these interacting, complex, complementary data systems to answer fundamental biological and medical questions. Dealing with them is nontrivial, because many questions we ask to answer from them fall into the category of computationally intractable problems, necessitating the development of heuristic methods for finding approximate solutions.

We develop methods for extracting new biomedical knowledge from the wiring patterns of systems-level, heterogeneous, networked biomedical data. Our methods link the patterns in molecular networks and the multi-scale network organization with biological function. In this way, we translate the information hidden in the wiring patterns into domain-specific knowledge. In addition, we introduce a versatile data fusion (integration) framework that can effectively integrate the information obtained from mining molecular networks with patient-specific somatic mutation data and drug chemical data to address key challenges in precision medicine: stratification of patients, prediction of driver genes in cancer, and re-purposing of approved drugs to particular patients and patient groups. Our new methods stem from novel network science approaches coupled with graph-regularized non-negative matrix tri-factorization, a machine learning technique for dimensionality reduction and co-clustering of heterogeneous datasets. We utilize our new framework to develop methodologies for performing other related tasks, including disease re-classification from modern, heterogeneous molecular level data, inferring new Gene Ontology relationships, and aligning multiple molecular networks.

Keywords:
data mining, new biomedical knowledge, disease re-classification

This work has been funded by ERC Starting and Consolidator grants, 2012-2023.

*Corresponding author, e-mail: natasa@cs.ucl.ac.uk

Beyond Patterns: Deep Understanding of Biology with Machine Learning

Predrag Radivojac

Indiana University, 700 North Woodlawn Avenue, Bloomington IN 47408, Indiana, USA

Abstract:
A major goal in computational biology is the development of algorithms, analysis techniques, and tools towards deep mechanistic understanding of life at a molecular level. In the process, computational biology must take advantage of the new developments in artificial intelligence and machine learning, and then move beyond pattern analysis to provide testable hypotheses for experimental scientists. This talk will focus on our contributions to this process and relevant related work. We will first discuss the development of machine learning techniques for partially observable domains such as molecular biology; in particular, methods for accurate estimation of frequency of occurrence of hard-to-measure and rare events. We will then show how these methods play key roles in inferring protein function and the phenotypic effect of coding sequence variants, with an emphasis on understanding the molecular mechanisms of human genetic disease. We further assessed the value of these methods in a wet lab where we tested the molecular mechanisms behind selected *de novo* mutations in a cohort of individuals with neurodevelopmental disorders. We finally discuss implications on future research in machine learning, genome interpretation, and precision health..

Keywords:
machine learning, computational biology, genome interpretation, precision health

*Corresponding author, e-mail: predrag@indiana.edu

Implications of dark proteome for precision medicine

Rost Burkhard

Technical University of Munich, Boltzmannstr. 3, Garching 85748, Germany

Abstract:
The objective of our group is to predict aspects of protein function and structure from sequence. The wealth of evolutionary information available through comparing the whole biodiversity of species makes such an ambitious goal achievable. Our particular niche is the combination of evolutionary information with machine learning. We developed methods to predict from sequence protein interactions (incl. networks), cellular localization, functional classifications and the effects of sequence variants upon molecular function and the organism. In this talk I will present the concept of the Dark Proteome and how protein disorder appears to play a unique role in evolution. Most of the talk will focus on the analysis of functional effects predicted between different people. Some surprising results appear relevant for precision medicine. A few examples will look into more detail by analyzing DNA/RNA/Protein-protein binding, membrane proteins and disorder.

Keywords:
protein function and structure prediction, machine learning, dark proteo

*Corresponding author, e-mail: assistant@rostlab.org

Identifying modifiers of somatic instability and age at onset in myotonic dystrophy type 1 by modeling genetic data

Dušanka Savić-Pavićević[1*], Jovan Pešović[1], Stojan Perić[2,3], Miloš Brkušanin[1], Goran Brajušković[1], Vidosava Rakočević-Stojanović[2,3]

[1] Center for Human Molecular Genetics, University of Belgrade-Faculty of Biology, Studentski trg 16, 11000 Belgrade, Serbia
[2] University of Belgrade-School of Medicine, Dr Subotića 8, 11000 Belgrade, Serbia
[3] Neurology Clinic, Clinical Center of Belgrade, Dr Subotića 6, 11000 Belgrade, Serbia

Abstract:
Myotonic dystrophy type 1 (DM1) is one of the most variable monogenic diseases, caused by expansion of CTG repeats in *DMPK* gene. Although the expansion size broadly correlates with DM1 severity, genotype-phenotype correlation is compromised by a pronounced somatic instability of mutated alleles. Somatic instability is biased towards further expansions throughout patient's life and influences severity of the disease. Modeling of somatic instability has shown that a large proportion of its variation is explained by the expansion size and the patient's age at sampling. The residual variation is attributed to factors determining individual-specific rate of somatic expansions, having a potential to modify DM1 phenotype.

Repeat-primed PCR screening of 242 patients from the Serbian registry for myotonic dystrophies revealed eight patients with repeat interruptions at the 3' end of expansions. Sanger sequencing showed that one patient had *de novo* CTC interruption, while others had various patterns of CCG interruptions: CTGCCG hexamers and CCG blocks mixed with individual CCG repeats. A striking feature of most patients was a later age at onset than expected for the corresponding expansion size. We further examined a potential modifying effect of interruptions on somatic instability and age at onset. Repeat-primed PCR demonstrated stable patterns of interruptions in blood cells over time. Somatic instability was quantified by small-pool PCR in eight patients with interrupted and four with uninterrupted expansions, resulting in dataset of app. 5000 sized alleles. To perform linear regression modelling of somatic instability and age at onset obtained dataset was combined with the published one, containing app. 20 000 alleles sized in more than 130 patients with uninterrupted expansions. Repeat interruptions were shown to statistically significantly contribute to less residual variations of somatic instability not accounted for by the expansion size and age at sampling, and to higher residual variations of age at onset explained by other factors modifying somatic instability than the expansion size and age at sampling. Our results established repeat interruptions as *cis*-acting and individual-specific genetic factors with a general stabilizing effect on somatic instability of *DMPK* expansions and positive modifying effect on age of DM1 onset, underlying their clinical importance for the disease course.

Keywords:
DNA repeats, genetic modifier, genetic instability, age at onset, modeling

This work has been funded by MESTD, Republic of Serbia (project 173016).

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

The Mechanisms of DNA Genetic Information Deactivation in Ion Cancer Therapy

Sergey N. Volkov

Bogolyubov Institute for Theoretical Physics, National Academy of Science, Metrologichna Str. 14-b, Kiev, Ukraine

Abstract:
A progressive method of the treatment of malignant tumors is radiation therapy. The mechanisms of action here are associated with the formation of reactive products of water radiolysis, which destroy the DNA strands. But under radiolysis another important process is the formation of long-lived molecular products. Among them a special interest causes the formation of hydrogen peroxide molecules (H_2O_2) that accumulate in the cell upon irradiation. These molecules can interfere with the realization of genetic information and play the important role the treatment of cancer.

The purpose of this work is to study the competitive binding of H_2O2 and H_2O molecules to the centers of specific and non-specific recognition of DNA macromolecule and to determine the possible effect of hydrogen peroxide on the mechanisms of the biological functioning of DNA

A computer analysis of the binding of water molecules and hydrogen peroxide to the centers of non-specific recognition of DNA double helix (main chain atoms) and specific recognition (nucleic acid atoms) is performed. Taking into account van der Waals-Coulomb interactions and hydrogen bonds, binding energies and structures of complexes of specific and non-specific recognition of the double helix of DNA with H_2O_2 and H_2O molecules are determined. The places of the most probable binding of hydrogen peroxide molecules to DNA are found. Conclusions are drawn about the possibility of blocking genetic activity of DNA macromolecule by hydrogen peroxide molecules in a cell.

The results of studying the complexation of hydrogen peroxide molecules with DNA recognition centers allow us to formulate a holistic mechanism for the effect of hydrogen peroxide on the processes of the transfer of genetic information in biological cells and can be used to improve the treatment of cancer and other chronic diseases.

Keywords:
DNA, cancer cells therapy, hydrogen peroxide, molecular modeling

This work has been funded by National Academy of Science of Ukraine

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

The New Wide Class of Symmetries in Long DNA-Texts. Elements of Quantum-Information Genetics

Sergey V. Petoukhov*, Vitaly I. Svirin

Mechanical Engineering Research Institute of Russian Academy of Sciences, Malyi Kharitonievsky pereulok 4, 101990, Moscow, Russia

Abstract:
Hidden symmetries in long sequences of oligonucleotides of single stranded DNA from their representative set are described. These symmetries are an addition to symmetries described by the second Chargaff's parity rule (%A ≅ %T and %G ≅ %C). These new symmetries and their rules concern collective probabilities of oligonucleotides from special tetra-alphabetical sets in long DNA-texts including all chromosomes of human and some model organisms. These rules of tetra-alphabetical probabilities were considered as candidacies for the role of universal rules of long DNA-sequences. On the basis of the known quantum-mechanic statement that quantum state of a multicomponent system is defined by the tensor product of quantum states of its subsystems, a quantum-informational model of genetic symmetries of these collective probabilities was proposed. In this model, nitrogenous bases C, T, G, A of DNA were represented as computational basis states of 2-qubit quantum systems. The biological meaning of the new quantum-information symmetries of long DNA-texts can be connected with the common ability of organisms to grow on the basis of incorporation into their body of new molecules of nutrients becoming new quantum-mechanic subsystems of the united quantum-mechanical organism.

Known binary-oppositional peculiarities of nitrogenous bases A, C, G, T allowed the algorithmic constructing square tables for interconnected alphabets of 4 monoplets, 16 doublets and 64 triplets with a strict order of these members in them. For the alphabets of 16 doublets and 64 triplets, we paid attention that the same strict order was generated by the tensor (Kronecker) exponentiation of the alphabetical matrix [C, A; T, G] in the second and third tensor powers. It testifies in favour that the mentioned tables can be considered as matrices and that the genetic system is related with the tensor product of matrices. Using known peculiarities of the degeneracy of the genetic code, the mentioned symbolic matrices can be algorithmically transformed into numeric matrices with their entries ±1 and with Walsh functions in their rows. We showed that these numeric genetic matrices are sums of sparse unitary matrices (unitary matrices are basis of calculations in quantum computers). Deep analogies exist between genetics and quantum computing.

Keywords:
symmetry, long DNA-sequences, probability, quantum informatics

This work has been funded by the budget of the Russian Academy of Sciences.

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

Use of phylogenetics in the study of viral genomes

Maja Stanojevic

University of Belgrade, Faculty of Medicine, Institute of Microbiology and Immunology

Abstract:
Rapidly evolving viruses comprise a number of new and emerging pathogens discovered over the last decades, many of them characterized by diverse pathogenic, biological, ecological features, such as human immunodeficiency virus (HIV) and hantaviruses. Research of the evolutionary biology of viruses provides an insight into the process of emergence of new agents, their adaptation and propagation in new hosts and may contribute to treatment and control of underlying diseases.

Molecular evolution involves structural changes in macromolecules induced in a process of selection and adaptation that are further manifested as resistance to antiviral drugs, immune escape mechanisms, viral tropism changes, diverse host-pathogen interactions etc. Study of molecular evolution provides an insight into the basis and extent of variability, allows the establishment of correlations between molecular phenomena and their phenotypic expression using diverse models of evolution within phylogenetics, phylodynamics, phylogeography. Research in the field of evolutionary biology provides invaluable new tools to understand the process of viral emergence, endemic or epidemic spread, as well as better understanding of the effects of therapy.

Keywords:
bioinformatics, DNA sequencing, viral evolution

This work has been funded by MESTD, Republic of Serbia (project 175024).

*Corresponding author, e-mail: maja.stanojevic@med.bg.ac.rs

Semantic Ion Vectors - deep learning applied to mass spectrometry

Starcevic A.[1], Zucko J.[1], Damir O.[1], Diminic J.[1] and Cindric M.[2]

[1] *Faculty of Food Technology and Biotechnology, University of Zagreb, Pierottijeva 6, 10000 Zagreb, Croatia*
[2] *Ruđer Bošković Institute, Bijenička cesta 54, 10000 Zagreb, Croatia*

Abstract:
Background: Protein mass spectrometry is the dominant method used for protein characterization. Peptide mass fingerprinting (PMF) is a phrase given to application of tandem mass spectrometry (MS2) to protein identification as the final goal of the method. In case of PMF, peptides do not fragment sequentially. To make things more complex, the process is not entirely random, with some fragmentations being more preferred over others. The resulting fragmentation spectrum captures fragment ions, which we observe as separate "peaks". Each peak is determined by a tuple of values: mass/charge (m/z) ratio (X-axis) and ion "intensity" (Y-axis), an underutilized and not fully understood value.

Results: A novel approach which relies on Deep learning techniques to capture distributed representations of peaks into Paragraph Vectors using unsupervised algorithm was employed to predict informative b and y ions and distinguish them from the "noise" peaks. Unlike Word embedding, a collective name used to describe the process of turning words and phrases into vectors of real numbers; this approach has taken a different turn. Numerical data (m/z and intensity couples) are turned into words (tokens), which were grouped into sentences and afterwards embedded using Paragraph Vectors. Several tokenization schemes have been implemented and performance of the resulting Ion Vectors have been tested under different parameters. Sole fitness criteria used was the performance of simple binary classification of peaks into "ions" (b and y) and "noise" (the rest). The best performing combination produced Semantic Ion Vectors, which were fed into classifier of choice and served as a model to make valuable predictions of "b-y ions" vs. "noise.

Conclusions: Resulting Semantic Ion Vectors can be used in variety of classification tools and provide accurate predictions of b-y ions. Tokenization method developed, efficiently reduced complexity of ion recognition task, by relying on constant "delta's" - distances between neighboring peaks rather than using m/z values directly. Intensity patterns and other information, commonly used as a set of rules in a synthetic manner (more suited to organic brain) were replaced by simple tokens and embedded into Paragraph Vectors using the vast number of spectral data in NIST repositories and neural networks.

Keywords:
bioinformatics, fragment spectrum, proteomics, b and y ions, Doc2Vec, deep learning, binary classification, big data, language processing, tokenization

This work has been funded by Croatian Science Foundation - Research project IP-06-2016 \"Exploring Gut Microbiome Equilibrium - MicroEquilibrium\".

*Corresponding author, e-mail: astar@pbf.hr

The rare layers of significance

Robert Šket[1], Tadej Debevec[2,3], Susanne Kublik[4], Michael Schloter[4], Anne Schoeller[4], Boštjan Murovec[5], Matevž Likar[6], Katarina Vogel Mikuš[6], Damjan Makuc[7], Klemen Pečnik[7], Janez Plavec[7], Igor Mekjavić[2], Ola Eiken[8], Zala Prevoršek[9], Marius Vital[10], Jenna Chandler[10], James M. Tiedje[10], Blaž Stres[1,11*]

[1] Group for Microbiology and Microbial Biotechnology, Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva ul. 101, 1000 Ljubljana, Slovenia
[2] Department of Automation, Biocybernetics and Robotics, Jozef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia
[3] Faculty of Sport, University of Ljubljana, Gortanova ul. 22, 1000 Ljubljana, Slovenia
[4] Research Unit for Comparative Microbiome Analysis, Helmholtz Zentrum München - German Research Center for Environmental Health,Ingolstädter Landstraße 1, D-85764 Neuherberg, Germany
[5] Machine Vision Laboratory, Faculty of Electrical Engineering, University of Ljubljana, Tržaška c. 25, 1000 Ljubljana, Slovenia
[6] Department of Biology, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva ul. 101, 1000 Ljubljana, Slovenia
[7] Slovenian NMR Centre, National Institute of Chemistry, Hajdrihova 19, 1000 Ljubljana, Slovenia
[8] Department of Environmental Physiology, Swedish Aerospace Physiology Centre, Royal Institute of Technology, Kungliga Tekniska Högskolan, SE-100 44 Stockholm, Sweden
[9] Group for Genetics, Animal Biotechnology and Immunology, Department of Animal Science, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva ul. 101, 1000 Ljubljana, Slovenia
[10] Michigan State University, Center for Microbial Ecology, 540 Plant and Soil Sciences Building, East Lansing, Michigan 48824-1325, USA
[11] Center for Clinical Neurophysiology, Faculty of Medicine, University of Ljubljana, Zaloška cesta 2, 1000 Ljubljana, Slovenia

Abstract:
We explored the assembly of intestinal microbiota in healthy male participants during the run-in (5 day) and experimental phases [21-day normoxic bed rest (NBR), hypoxic bedrest (HBR)], and hypoxic ambulation (HAmb) in a strictly controlled laboratory environment, balanced fluid, and dietary intakes, controlled circadian rhythm, microbial ambiental burden, exercise and 24/7 medical surveillance. The fraction of inspired O2 and partial pressure of inspired O2 were 0.209 and 133.1 ± 0.3 mmHg for NBR and 0.141 ± 0.004 and 90.0 ± 0.4 mmHg for both hypoxic variants (HBR and HAmb; ~4,000 m simulated altitude), respectively. A number of parameters linked to intestinal transit spanning Bristol Stool Scale, defecation rates, zonulin, α1-antitrypsin, eosinophil derived neurotoxin, bile acids, reducing sugars, short chain fatty acids, total soluble organic carbon, water content, diet composition, and food intake were measured. In addition, intestinal electrical conductivity, sterol and polyphenol content and diversity, indole, aromaticity and spectral characteristics of dissolved organic matter next to 1H- and 13C-NMR metabolomes and metals X-ray fluorescence were assessed. A new in-house database was established to integrate all measured variables on host physiology, diet, experiment, immune and metabolic markers (n > 350).

The abundance, structure, and diversity of butyrate producing microbial community were assessed using the two primary bacterial butyrate synthesis pathways, butyryl-CoA: acetate CoA-transferase and butyrate kinase genes. The structure and diversity of bacterial microbial community was assessed using 16S rRNA amplicon sequencing. Shotgun metagenomes were analyzed at various taxonomic and functional levels.

Bayesian network analysis was used to derive the first hierarchical model of initial inactivity mediated deconditioning steps over time. The PlanHab wash-out period corresponded to a profound life-style change (i.e., reintroduction of exercise) that resulted in stepwise amelioration of the negative physiological symptoms, indicating that exercise apparently prevented the crosstalk

between the microbial physiology, mucin degradation and proinflammatory immune activities in the host, carrying much resemblance to mammalian hibrernators and yo-yo dieting.

Keywords:
bioinformatics, data mining, computer science, nucleic acids, sequencing

This work has been funded by the European Union programme FP7 (PlanHab project; Grant no. 284438; https://ec.europa.eu/research/fp7; http://planhab.com), the European Space Agency (ESA) Programme for European Cooperating States (ESTEC/Contract No. 40001043721/11/NL/KML: Planetary Habitat Simulation; www.esa.int/), and the Slovene Research Agency (Contract no. L3-3654: Zero and reduced gravity simulation: The effect on the cardiovascular and musculoskeletal systems; https://www.arrs.gov.si/). BS, ZP, RS, BM, ML acknowledge the support of Slovenian Research Agency (https://www.arrs.gov.si/) next generation sequencing research grants and Young Research Fellowship to RS (SRA#37426). RS and BS acknowledge the support of National Scholarship Programme (Planetary Emergencies: Medicine and Biotechnology) granted by the World Federation of Scientists in 2017/2018 (http://www.federationofscientists.org/). Also, this work was has been in part funded by SPADE project proposal within H2020: Research & Innovation Actions (RIA) call Encouraging the re-use of research data generated by publically funded research projects.

*Corresponding author, e-mail: blaz.stres@bf.uni-lj.si

Circular codes in the genetic information

Lutz Strüngmann[1*], Elena Fimmel[1], Christian Michel[2], Francois Pirot[2] and Jean-Sébastien Sereni[2]

[1] *Institute for Mathematical Biology, Faculty of Computer Science, Mannheim University of Applied Sciences, Paul-Wittsack-Str. 10 68163 Mannheim, Germany*
[2] *Theoretical bioinformatics, ICube, University of Strasbourg, C.N.R.S., 300 Boulevard Sébastien Brant 67400 Illkirch, France*

Abstract:
By an extensive statistical analysis in genes of bacteria, archaea, eukaryotes, plasmids and viruses, a maximal C3 self-complementary trinucleotide circular code was found to have the highest average occurrence in the correct reading frame of the ribosome during translation, when compared to the two shifted frames. Moreover, dinucleotide circular codes have been observed in non-coding regions of eukaryotic genomes while special tetranucleotide codes were hypothesised as one of the ancestral codes during evolution and their occurrences seem to be related to phylogenetics.

The circular codes are assumed to play an essential role in maintaining the reading frame during the translational process in the ribosome. A theoretical framework has therefore been developed over the last years using several methods and theories from mathematics, bioinformatics and computer sciences. In this talk we consider various aspects of circular codes, i.e. circular codes that consist of n-nucleotide words over the alphabet of nucleic bases {A,C,G,T} for different n. Based on recent results we discuss basic properties of such circular codes, give construction methods for maximal circular codes and put our work in an evolutionary context. In particular, we will consider mixed circular code which could have been an intermediate code during evolution when passing from the dinucelotide world to the trinucleotide world.

Keywords:
circular codes, translation, genetic code

*Corresponding author, e-mail: l.struengmann@hs-mannheim.de

Comparative genomics of a recently emerging epidemic on banana in Sub-Saharan Africa

David J. Studholme

Biosciences, University of Exeter, Stocker Road, Exeter, EX4 4QD, United Kingdom

Abstract:
Banana *Xanthomonas* wilt (BXW) is a disease on banana crops that emerged in east Africa in the 21st century. It is currently devastating the livelihoods of smallholder farmers in Uganda and neighboring countries. The causative agent is a bacterial strain called *Xanthomonas campestris* pv. *musacearum* (Xcm). This same pathogen, *Xcm*, is also responsible for endemic wilt disease of enset, which is a crop plant of great importance to food security in Ethiopia. Furthermore, *Xcm* is closely related to a bacterial strain recently emerging as a pathogen on maize in the USA.

We sequenced genomes of multiple isolates of the *Xcm* pathogen from a range of times and geographical locations in an attempt to gain insight into the emergence of the epidemic on banana and its evolutionary trajectory of its emergence by comparison with genomes of related bacteria. We identified specific candidate virulence and avirulence factors that are associated with adaptation to banana and/or enset. We find evidence that the east African epidemic probably originates from two distinct introductions of the pathogen from Ethiopia. Comparative genomics has also informed the development of molecular diagnostic tools for detection and identification.

To complement work on pathogen genomics, a parallel programme of genome sequencing has been applied to the host organisms, *i.e.* banana and enset, revealing significant genetic variation that could ultimately be used for crop improvement. The work presented in this overview includes contributions from many collaborators from several institutions in Europe and Africa.

Keywords:
genomics, DNA sequencing, bacteria, crops, detection

This work has been funded by The European Community Horizon 2020 grant Project ID 727624, "Microbial uptakes for sustainable management of major banana pests and diseases (MUSA)".

*Corresponding author, e-mail: d.j.studholme@exeter.ac.uk

Why Human Brain Networks are Hyperbolic?

Bosiljka Tadić[1*] and Miroslav Andjelković[2]

[1] *Department of Theoretical Physics, Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia*
[2] *Vinca Institute of Nuclear Sciences, University of Belgrade, 12-14 Mike Petrovića Street, 11351 Vinca, Belgrade, Serbia*

Abstract:
Recently, the study of graphs representing various complex systems has been extended beyond the standard graph-theoretic measures. The use of methods of algebraic topology enabled revealing the higher organized structures and related hidden geometries that can appear in the graph. More specifically, the Q-analysis based on the algebraic topology of graphs identifies elementary geometric shapes or simplexes (triangles, tetrahedrons and higher order cliques) and how they connect to make more substantial structures or simplicial complexes. The original composition of these basic geometry descriptors is unique for a particular network; furthermore, it can induce emergent hyperbolicity or *negative curvature*, a measure of nodes proximity in the graph-metric space, which often associates with an improved function of the network. However, how the hyperbolic geometry evolves [T3] to support the network's capacity remains a challenging issue, depending on the nature of the complex system in question. In this lecture, we discuss such hidden structures in conjunction with the functional properties of different types of human brain networks. More specifically, we consider the network structures that are originating from the aggregated fNMR imaging data recently described in [C1,C2], and the graphs mapping the brain activity patterns during social communications recorded by EEG [E1,T1,T2]. Using the generalized 4-point Gromov hyperbolicity criterion for graphs, we demonstrate that these brain networks are hyperbolic, and by performing Q-analysis, we determine the structure of underlying simplicial complexes in them. By comparing brain networks obtained from selected sets of data in conjunction with various resolution, female/male participants, and the level of cross-brain coordination, we attempt to find out how these hidden geometries vary, which may suggest the potential neuro-dynamical origin of the observed negative curvature.

Keywords:
brain imaging data, brain networks, simplicial complexes, gromov hyperbolicity

This work has been funded by The Slovenian Research Agency (ARRS).

*Corresponding author, e-mail: Bosiljka.Tadic@ijs.si

On the representation of continuous spaces in the mammalian brain

Alessandro Treves[1,2*], Davide Spalla[1], Eugenio Urdapilleta[3], Federico Stella[4] and Zeynep Kaya[1]

[1] *Cognitive Neuroscience, SISSA, via Bonomea 265, 34136 Trieste, Italy*
[2] *Center for Neural Computation, NTNU, Olav Kyrres gate 9, 7030 Trondheim, Norway*
[3] *Instituto Balseiro and CONICET, Av E. Bustillo, R8402AGP San Carlos de Bariloche, Río Negro, Argentina*
[4] *IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria*

Abstract:
I will report on work in progress that addresses the issues of whether metric representations of any continuous space can be set up in the brain, and whether multiple such representations can coexist, from three different directions.

First, mathematical analyses of a model of planar representations by grid units, with Davide Spalla, have indicated that orthogonal grid representations can indeed coexist, within a capacity limit; so that, for example in rodents, but possibly also in bats and in other species, the possibility is there that different grids (of the type discovered by Hafting et al, 2005, in the Moser lab) may be expressed by the same cells embedded within the same network, at least in theory.

Second, simulations of grid formation on a sphere (similar to those published by Stella et al, 2013), run by Federico Stella and Eugenio Urdapilleta, have shown a powerful clustering effect – such that grid units that have clustered together in one metric representation may never find the strength, so to speak, to disband and set up an orthogonal metric.

Third, experiments yet to be run on proficient adult human bilinguals, a follow up on others currently run by Zeynep Kaya on ERP responses to the perception of vowels, test whether distinct representations can keep the different metrics in vowel space that we see, across subjects, in monolinguals. What shall we find? Preliminary results may be presented if obtained before the meeting.

Keywords:
neural computation, grid cells, storage capacity, curved geometry, vowels

This work has been funded by Human Frontier Science Program RGP057/2016.

*Corresponding author, e-mail: ale@sissa.it

Combining structural aggregation propensity and stability predictions to re-design protein solubility

Marcos Gil-Garcia[1], Nathalia Varejão[1], Michal Jamroz[2], Aleksander Kuriata[2], Marta Diaz-Caballero[1], Jara Lascorz[1], Susanna Navarro[1], David Reverter[1], Sebastian Kmiecik[2] and Salvador Ventura[1*]

[1] Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Cerdanyola del Vallès, Bellaterra (Barcelona), Spain
[2] Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Pasteura 1, PL-02-093 Warsaw, Poland

Abstract:
The aggregation propensity of each particular protein seems to be shaped by evolution according to its natural abundance in the cell. The production and downstream processing of recombinant polypeptides implies attaining concentrations that are orders of magnitude above their natural levels, often resulting in their aggregation; a phenomenon that precludes the marketing of many globular proteins for biomedical or biotechnological applications. Therefore, there is a huge interest in methods aimed to rise proteins solubility above their natural limits. Here, we demonstrate that an updated version of our AGGRESCAN 3D structural aggregation predictor, that now takes into account protein stability, allows designing mutations at specific positions in the structure that improve the solubility of proteins without compromising their conformation. Using this approach, we have designed a highly soluble variant of the Green Fluorescent Protein (GFP) and a human single-domain VH antibody displaying significantly reduced aggregation propensity. Overall, our data indicate that the solubility of unrelated proteins can be easily tuned by in silico-designed non-destabilizing amino acid changes at their surfaces.

Keywords:
protein aggregation, protein stability, protein structure, green fluorescent protein, Aβ peptide, single-domain antibodies

*Corresponding author, e-mail: salvador.ventura@uab.es

Inference and prediction in molecular biological systems

Jose M. G. Vilar[1, 2*]

[1] *Biofisika Institute (CSIC, UPV/EHU) and Department of Biochemistry and Molecular Biology, University of the Basque Country UPV/EHU, 48080 Bilbao, Spain*
[2] *KERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain*

Abstract:
The traditional computational systems approach seeks to provide the emergent collective behavior from the properties of the constituent elements. The molecular and mechanistic complexity of typical biological systems, however, often results in a nonexistent, deficient, or misleading characterization of the underlying molecular components and their interactions. Here, I discuss the state of the art of novel approaches that function in reverse, namely, that use the emergent behavior to infer precise properties of the molecular components and that are able to make accurate predictions about the system behavior even without a basic molecular description. I illustrate the main points with key examples in gene regulation, protein aggregation, and odor perception.

Keywords:
protein aggregation, gene regulation, molecular modeling, inverse methods

*Corresponding author, e-mail: j.vilar@ikerbasque.org

Predicting Disease from Gut Microbiota Codon Usage Profiles

Maja Kuzman[1], Anamaria Elek[1], and Kristian Vlahoviček[1*]

[1] *Bioinformatics group, Division of Biology, Department of Molecular Biology, Faculty of Science, University of Zagreb, Horvatovac 102a, 10000 Zagreb, Croatia*

Abstract:
Metagenomics projects use next-generation sequencing to unravel genetic potential in microbial communities from a wealth of environmental niches, including those associated with human body and relevant to human health. In order to understand large datasets collected in metagenomics surveys and interpret them in context of how a community metabolism as a whole adapts and interacts with the environment, it is necessary to extend beyond the conventional approaches of decomposing metagenomes into microbial species' constituents and performing analysis on separate components. By applying concepts of translational optimization through codon usage adaptation on entire metagenomic datasets, we demonstrate that a bias in codon usage present throughout the entire microbial community can be used as a powerful analytical tool to predict for community lifestyle-specific metabolism. Here we demonstrate this approach combined with machine learning, to classify human gut microbiome samples according to the pathological condition diagnosed in the human host.

Keywords:
codon usage, translational optimization, machine learning, metagenomics, microbial community, gut microbiota

*Corresponding author, e-mail: kristian@bioinfo.hr

Bioinformatics strategy for rare disease diagnostics in the era of next-generation sequencing

Branka Zukic[1*] and Sonja Pavlovic[1]

[1] *Laboratory for Molecular Biomedicine, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11010 Belgrade, Serbia*

Abstract:
We are the first institution is Serbia that applied next-generation sequencing (NGS) methodology and biological big data analysis in research and diagnostics of rare diseases. Over 100 rare disease patients were analyzed using Clinical-Exome Sequencing TruSight One Gene Panel (Illumina, San Diego, CA, USA) which includes 4813 disease-associated genes described in the OMIM database until 2013. Pipeline (or a "primary analysis") and secondary analysis which generated FastQ, BAM and VCF files were performed by the Illumina MiSeq instrument and MiSeq Reporter software using default settings. Variant calling files were further annotated and examined using the Illumina VariantStudio 2.2 Data Analysis Software (Illumina, San Diego, CA, USA). For genetic diseases, filtration and prioritization of variants were performed according to "in-house" pipeline. Variants were analyzed by various *in silico* softwares including PolyPhen2, SIFT and Mutation Taster and classified according to ACMG guidelines. Variants and their association with clinical symptoms of the patients were checked if previously described in HGMD, OMIM, ClinVar and Varsome databases. Variants selected by these criteria were confirmed by conventional Sanger sequencing and parents' samples were analyzed whenever available.

Clinical exome sequencing enabled diagnosis of more than 50 different diagnosis (hematological, metabolic, endocrinological, pulmonary, immunological, orthopedic, dermatological, ophthalmological, epileptic encephalopaties etc). It was particularly important for genetically heterogeneous diseases, such as glycogen storage diseases, branched-chain organic acidurias, primary ciliary dyskinesia, MODY or mitochondriopathies. Moreover, different diseases with overlapping clinical manifestations were accurately diagnosed. Also, we used TruSeq-Amplicon Cancer Panel to analyse different childhood and adult rare hematological malignancies. Besides studying diagnostic and prognostic malignancy markers, we designed "in-house" virtual pharamocogenomic panel, and performed association studies of pharmacogenomic markers and the course and outcome of rare hematological malignancies, resulting in recommendations for therapeutic modalities in accordance with genomic profile of the patient.

Our approach leads to timely and accurate genetic testing, sets definite diagnosis and enables rapid implementation of optimal therapy for patients with rare diseases in Serbia. Furthermore, functional characterization studies contribute to unambiguous diagnostic interpretation of novel genetic variants worldwide.

Keywords:
next-generation sequencing, TruSight One Gene Panel, rare diseases, biological big data analysis

This work has been funded by MESTD, Republic of Serbia (project III41004).

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

ORAL PRESENTATIONS

Prediction of Pharmacological Treatment for Patients with Coronary Artery Disease

Bojana Andjelkovic Cirkovic[1,2*], Nenad Filipovic[1,2]

[1] *Faculty of Engineering, University of Kragujevac, Sestre Janjic 6, 34000 Kragujevac, Serbia*
[2] *BioIRC, Research and Development Center for Bioengineering, Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia*

Abstract:
Optimal medical therapy for patients with coronary artery disease (CAD) is necessary to prevent disease progression and recurrent cardiovascular events. Due to high morbidity and mortality of patients with CAD, the development of reliable predictive tools that could help clinicians to tailor patient-specific therapy modulation is recognized as public health benefit.

Several groups of medicaments are essential in therapy, but to reduce the risk of CAD complications as well as medication side effects, it is very important to include all relevant parameters in carefully designed manner. In this study, we propose a data mining approach for the development of predictive decision-making system for CAD medication choice.

A multi-label classification model was developed based on class imbalanced patients' electronic health records which consisted of numerous predictor variables, ranging from clinical, demographic, lifestyle, blood test, cardiovascular disease risk factors data, etc. The information about medicaments that have already been taken was also included. Real-world data from 118 patients and 13 output labels were exploited by data mining techniques and it has been found that demographic, clinical, risk factors and biohumoral markers data in conjunction with information about current therapy and disease progress form the most informative predictive base. The recommended multi-label predictive model was developed by employing the Ensemles of Classifiers Chain and Decision Tree algorithms. The model demonstrated the stable performance in every single labeled classification and scored 81.1% for *Accuracy* and 0.051 for *Hamming Loss*.

Keywords:
multi-label classification, coronary artery disease, pharmacological treatment

This work has been funded by EC HORIZON2020 689068 SMARTool, and MESTD, Republic of Serbia (projects III41007 and OI174028).

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

FTO haplotyping underlines high obesity risk for European populations

Roman Babenko

The Federal Research Center Institute of Cytology and Genetics, Prospekt Lavrentyeva 10, 630090 Novosibirsk, Russian Federation

Abstract:
We analyzed the population specific haplotype profiles of the FTO gene genomic locus hotspot identified by Genome Wide Association Studies (GWAS) for the high obesity risk by scrutinizing eighteen 1000G populations from 4 continental groups. The hotspot is located in FTO gene intron 1 spanning around 40kb. We reconstructed the ancestral state of the locus, which comprised 'healthy' major allele found in all populations, and two minor 'risky' alleles, each one specific for African and European populations, correspondingly. The allele locus structure and frequency distribution underscores the high risk allele specifically for European population. South Asian populations take the second place on the 'risky' allele frequency, while East Asian populations have the minimal ratio of risky allele. African populations specific allele was only 'partially' risky, while the majority of GWAS SNPs were manifested by healthy alleles'. These observations corroborate the previous reports on the FTO locus implication in population specific manner as well as WHO BMI index population distribution. Thus, the conclusions presented imply FTO locus analyzed is rather a major genetic determinant of the genetic obesity risk from the GWAS SNPs set.

Keywords:

*Corresponding author, e-mail: babe-roman@yandex.ru

Effect of mutant NMDA receptors on oscillations in a model of hippocampus

A. S. Batova, A. N. Bugay* and E. B. Dushanov

Joint Institute for Nuclear Research, Joliot-Curie 6, 141980 Dubna, Russia

Abstract:
Cognitive functions of the brain (attention focusing, memory, spatial orientation and executive functions) are considered in the context of the occurrence in the brain of certain rhythms (or waves) characteristic for each of these functions. Hippocampus is a part of the brain responsible for short-term and spatial memory, as well as participating in memory consolidation. To a large extent, the hippocampus contributes to the generation θ and γ rhythms. In the development of various kinds of neurodegenerative diseases, as well as after exposure to ionizing radiation, a number of disorders are associated with changes in the synaptic transmission. It is suggested that these effects are of a genetic nature and are associated with the occurrence of mutations in genes encoding key proteins. The aim of this work was to develop a computational approach capable to establish a link between mutations in a genetic structures and high-level observable properties of neural tissue.

In the course of the work, molecular-dynamic modeling of NMDA receptor ion channel gating was carried out. The receptor was constructed from mutant forms of proteins, The analysis of the resulting configurations made it possible to determine the change in the conductivity of the ion channels, and the obtained data were implemented in the structure of model neural network of hippocampus. For each segment of interneurons and pyramidal neurons in the neural network, the corresponding ion channels and synaptic contacts were taken into account in detail. Based on this model, the calculation of the onset of synchronous neuronal oscillations contributing to θ and γ rhythms was carried out.

As a result, it was possible to identify the effect of single and double point mutations in the protein units NR2 of NMDA receptors on the generation of θ and γ rhythms by the neural network. Obtained results can be adapted for analysis and evaluation of possible cognitive impairments arising from the action of radiation and other negative factors.

Keywords:
neural network, molecular dynamics, neurodegeneration

*Corresponding author, e-mail: bugay_aleksandr@mail.ru

Clustering and classification of human microbiome data: evaluating impact of different settings in bioinformatics workflows

Sanja Brdar[1*], Tatjana Lončar-Turukalo[2], Vladimir Crnojević[1] and Blaž Stres[3]

[1] *BioSense Institute, University of Novi Sad, Dr Zorana Đinđića 1, 21000 Novi Sad, Serbia*
[2] *Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia*
[3] *Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, SI-1000 Ljubljana, Slovenia*

Abstract:
Microbiome studies contain immense potential for innovations in medicine and agriculture when coupled to multiscale data on host health and physiology. For efficient interrogation of microbiome data with multiscale data robust computational and modelling approaches are needed. Numerous processing (quality filtering, picking operational taxonomic unit (OTU), creating OTU table) and final clustering or classification steps are utilized for providing valuable insights. In this study we explored and quantified changes in the results of clustering and classification, with respect to different settings in QIIME bioinformatics workflow.

The data from Moving pictures of human microbiome study were used encompassing millions of 16S rRNA gene sequences from 1967 samples, obtained from three body sites (oral, skin, gut) of male and female subject over time of 15 and 6 months, respectively. Variability of the results was explored in the light of two different reference databases (Greengenes and Silva), and four level of taxonomy induced by different similarity thresholds (90 to 99%). These settings highly impacted the number of features (6637 to 87069) and values in the OTU tables (i.e. matrix sparseness). These tables were input for classification tasks, while prior to the clustering beta diversities were measured from OTU tables and used as pairwise distance matrices.

Several classification and clustering algorithms were tested and performance of the best - Random Forest classifier (RF) and Spectral clustering (SC) is presented. RF demonstrated capacity to learn on sparse data and provided almost the same results in classifying samples regardless of reference database and settings. Classification accuracy, estimated by 10-fold cross-validation, was at the level of 96%. Efficient classification of healthy baseline microbiome data provides grounds for robust modelling of microbiome data with health related data sets. SC showed large fraction of variability on measured beta-diversity matrices and high sensitivity to any change in the parameters. Agreement of the clustering results with true classes measured by adjusted rand index varied from 0.35 to 0.59, with mean value 0.51±0.06 standard deviation. Classification better separated samples according to their labels than clustering; however as clustering is prevailing and no labels are available in many microbiome studies, further improvements of clustering by dimensionality reduction, ensemble or semi-supervised approaches should be explored.

Keywords:
bioinformatics, microbiome, 16S rRNA, OTU table, machine learning

*Corresponding author, e-mail: brdars@uns.ac.rs

3D reconstruction of cerebral reactive oxygen species formation as a gold standard for future *in vivo* molecular imaging approaches

Ana I Casas[1], Anne-Sophie Herard[2], Thierry Delzescaux[2], Javier Egea[3], Harald HHW Schmidt[1]*

[1]*Department of Pharmacology & Personalised Medicine. Faculty of Health, Medicine & Life Sciences. Maastricht University. Universiteitssingel 50, 6229 ER, Maastricht, The Netherlands*
[2] *Commissariat à l'Energie Atomique (CEA) - Molecular Imaging Research Center (MIRCen), 92265 Fontenay-Aux-Roses, France*
[3]*Instituto de Investigación Sanitaria, Servicio de Farmacología Clínica, Hospital Universitario de la Princesa. Facultad de Medicina. Universidad Autonoma de Madrid. Madrid, Spain*

Abstract:
Background: Ischemic stroke is now considered the second leading cause of death and the first one of disability in industrialized countries. Reactive oxygen species (ROS) resulted to be key players in the stroke pathomechanism. In fact, NADPH oxidase 4 and 5 (NOX4/5), main physiological ROS sources, have been recently identified as promising therapeutic targets for brain ischemia. Preliminary both *in vitro* and *in vivo* animal studies suggest that direct enzymatic blockage of NOX4/5 using pharmacological treatments lead to a significant improvement of the neuro-motor outcome, reduction of infarct size, less neuronal death and direct neuroprotection.

Objectives: So far, almost no *in vivo* ROS imaging has been performed in stroke although we strongly believe that only 3D-imaging technologies will enable us to visualize ROS released at specific time-points post-stroke. During this STSM we investigated the specific time-frame of NOX4/5-dependent ROS formation. We established the imaging method needed for ROS detection in complete brain for future high definition 3D reconstruction. Understanding the precise ROS kinetics post-stroke will allow us to define the best therapeutic window. Then, optimal treatment at optimal time could be directly translated to the clinics. Therefore, we will use high precision imaging for fine-tuning the most suitable timing for treatment Finally, both the treatment strategy and the imaging approach will be subsequently translated into the clinics as part of a phase II clinical trial which is currently under preparation.

Results & Methods: During this STSM we focused on the first experimental stage using a stroke animal model of brain ischemia followed by different pharmacological approaches or transgenic mice. As a final step, we established an *ex vivo* molecular imaging technique focused on multi-slicing and staining of stroked brains which in future steps will be used for 3D reconstruction brain images. Basic multiscale imaging scanning and digitalisation was conducted while establishing the ex vivo molecular imaging technique. In later stages of the project we will use high definition multi scale modelling for 3D brain reconstruction.

Conclusion: Pre-clinical imaging techniques are considered a key step in drug development and patient stratification. Towards the clinics, suitable dosage and treatment time-frame provide a direct prediction of treatment response and patient prognosis. Mid-/long-term outcome of stroke patients critically depends on choosing the most suitable timing of treatment. Thus, stroke *in vivo* imaging could lead to a direct improvement of brain ischemia treatment.

Keywords:
3D reconstruction, imaging, stroke, reactive oxygen species, NADPH oxidase 4/5, treatment, clinics.

*Corresponding author, e-mail: a.casasguijarro@maastrichtuniversity.nl
h.schmidt@maastrichtuniversity.nl

The role of mRNA secondary structure in the control of translation and mRNA degradation in E. coli

Z. Chervontseva[1,2*], E. Khodzhaeva[3], I. Ponamareva[4] , M.S. Gelfand[2,4,5], A.A. Mironov[1,2,4,5]

[1] *Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, 1-73 Leninskiye Gory, Moscow, Russian Federation*
[2] *Kharkevich Institute for Information Transmission Problems RAS, 19-1 Bolshoy Karetny per., Moscow, Russian Federation*
[3] *Lomonosov Moscow State University, Faculty of Biology, 1-12 Leninskie Gory, Moscow, Russian Federation*
[4] *Higher School of Economics, Faculty of Computer Science, 3 Kochnovsky Proezd, Moscow, Russian Federation*
[5] *Skolkovo Institute of Science and Technology, 3 Nobel, Moscow, Russian Federation*

Abstract:
Protein biosynthesis is an energy-consuming process, and hence it is tightly regulated at several levels. One mechanism of regulation is mRNA secondary structure. We have used published genome-wide DMS-Seq data on mRNA secondary structure, ribosomal profiling determining translation efficiency, data on mRNA degradation, ribosome drop-off rates calculated using the ribo-seq data, supplemented by mutation data calculated by comparison of 350 completely sequenced E. coli strains to study how the mRNA structure affects gene expression, and how it evolves.

Comparison of the DMS-Seq and ribosomal profiling data has confirmed published observations that the translation efficiency is lower in more structured mRNAs. On the other hand, the ribosomal drop-off rates are lower in more structured regions. One possible explanation for this paradox could be that the ribosomes do not interact with each other in less populated regions, translated at a slower rate.

Expression of protein subunits in equimolar complexes needs to be tightly coordinated. As expected, the mass-spec data shows that the concentrations of such proteins are highly correlated. If two subunits are encoded in one operon, the translation rates of the respective genes should be similar. Indeed, both the translation rates and the level of secondary structure are more correlated in same-operon genes encoding equimolar subunits, compared with control same-operon genes.

It has been suggested that the terminator hairpin protects mRNA from degradation. While we have not observed overall correlation between the level of secondary structure and degradation rates of mRNA, relatively more structured regions of individual mRNAs persist longer that less structured regions. At that, efficiently translated mRNAs tend to live longer, likely due to the protective role of translating ribosomes.

Analysis of mutation rates demonstrates that, contrary to expectations, nucleotides forming base-pairs tend to be mutated more often. Mutations tend to occur in certain contexts, but Monte-Carlo simulations show that this may be due to the fact that such contexts preferentially form secondary structures.

Overall, integration of diverse experimental data supports the hypothesis that the mRNA secondary structure has a functional role beyond known regulatory elements and influences evolution of mRNA sequences.

Keywords:
bioinformatics, mRNA secondary structure, translation rates, evolution.

This work has been funded by Russian Science Foundation (grant 18-14-00358).

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

DiNGO: stand-alone application for GO and HPO term enrichment

Radoslav Davidović[1], Branislava Gemović[1], Nevena Veljković[1], Vladimir Perović[1*]

[1]*Centre for Multidisciplinary Research and Engineering, Vinca Institute of Nuclear Sciences, University of Belgrade, Mike Petrovića-Alasa 12-14, 11001 Belgrade, Serbia*

Abstract:
DiNGO is a software application which was developed based on open source code from BiNGO (The Biological Networks Gene Ontology tool) a Java-based tool aimed to determine which Gene Ontology (GO) categories are statistically overrepresented in a set of genes or a subgraph of a biological network. BiNGO is widely used as a Cytoscape plug-in which brings advantages through GUI, but also restricts its application only on a single ontology and affects its processing speed. These limitations were motivating factors for creating DiNGO. We implemented DiNGO as a stand-alone application, that offers increased processing capability and allow user to perform enrichment analysis of other ontologies.

DiNGO uses same set of parameters as BiNGO, but in contrast to latter DiNGO receives user input from command line interface instead of Cytoscape GUI, and prints results directly in file. Such implementation decision was taken to improve processing speed. Due to increasing interests, among scientists and clinicians for the application of Human Phenotype Ontology (HPO) in clinical settings, we decided to exemplified DiNGO usage by HPO terms enrichment. The application uses HPO ontology file in obo format by default. For user convenience DiNGO already contains HPO annotation file formatted as BiNGO default file format. However, DiNGO allows for utilization of others file formats enabled by BiNGO.

In order to test processing speed of BiNGO and DiNGO we used input file that contained 1000 clusters, each cluster contained in average 1244 proteins. Our analysis showed that DiNGO performed this task 2 times faster than BiNGO. Next, we provided the illustrative example of DiNGO application on 35 consensus cancer genes and 638 genes annotated with HPO term HP 0002664:neoplasm.

In conclusion, DiNGO is computationally efficient modification of popular enrichment program BiNGO which allows for implementation on ontologies different from GO.

Keywords:
BiNGO, DiNGO, human phenotype ontology, enrichment

This work has been funded by MESTD, Republic of Serbia (projects ON173001, ON173049).

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

Characterization of novel risk predictors in colorectal cancer

Seçil Demirkol[1], Andreas Lindler[2], Steven Carberry[2], Ali Osmay Güre[1] and Jochen Prehn[2]

[1] *Molecular Biology and Genetics Department, Bilkent University, Ankara, Turkey*
[2] *Physiology and Medical Physics Department, Royal College of Surgeons, Dublin, Ireland*

Abstract:
Aims: Prognostic biomarkers for cancer have the power to change the course of disease if they add value beyond known prognostic factors, if they can help shape treatment protocols, and if they are reliable. Here we aimed to biologically understand and characterize two previously published prognostic signatures further at proteomic and mRNA levels.

Methods: In colaboration with Prof. Jochen Prehn in RCSI, Ireland, we performed protein arrays (PA) of primary tumors from 48 Turkish colon cancer patients. *In silico* we worked on a mathematical model of BCL-2 protein interactions (DR_MOMP) which was established by Prehn group previously, and investigated the clinical/biological parameters which can contribute to DRMOMP in terms of prognostic prediction utilizing publically available gene expression data. For this purpose, TCGA RNAseq and related clinical data were analyzed and compherehensive bioinformatic analyses were performed using previously known subtypes of colon cancer, clinical confounding factors, biological processes such as EMT, apoptosis, stemness and immune involvement for stratification.

Results: We previously identified two novel mRNA based biomarkers, ULBP2 and SEMA5A, which are associated with overall and recurrence-free survival independent of all confounding factors in colon cancer *in silico* and *ex vivo*. Our findings also showed that the prognostically distinct subgroups identified by these biomarkers are also biologically different at the RNA level *in silico* in terms of EMT status and inflammatory gene expression. Utilizing the PA output of primary colon tumors which we have not received yet, we further plan to work on protein level expression differences primarily for apoptosis related genes between the prognostic groups.

Our bioinformatic analysis indicated that DRMOMP can prognostically stratify patients with high Oncotype Dx risk score, with CMS 1&3 tumors, high ALMAC risk score, more proliferative and less stem cell-like gene expression.

Keywords:
bioinformatics, TCGA, colon cancer, prognosis

*Corresponding author, e-mail: secil.demirkol@bilkent.edu.tr

Reduction method for reaction-diffusion equations from biology

Aurelia Florian, Jelena Ivic, Alina Pauna*, Mihai Stoicescu

University of Craiova, Department of Physics, 13, A.I. Cuza, Craiova, Romania

Abstract:
Many dynamical phenomena in biological systems are modeled through nonlinear partial differential equations. There is not a standard procedure for solving such equations, but various approaches allow to establish if an equation is integrable and, in the positive case, to find classes of its solutions. This paper will present two such approaches, both of them based on the idea of simplifying the studied equations by a reduction procedure. The first approach proposes a reduction by attaching a flow-type equation to the initial equation. By combining the two, a "reduced" equation for the attached flux, simpler to be investigated, will be generated. The second approach will refer to the similarity reduction used in the study of the classical symmetry for the partial differential equations. By this reduction, the problem of finding solutions for complicated equations is replaced with that of solving a "determining" system. Both approaches will be illustrated on equations describing dynamical phenomena in biological systems.

The reduction method of the initial equations based on attached flow-type equations will be applied to an equation modeling the propagation of nerve pulses through neurons. It is a generalized Boussinesq equation of the form:

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial}{\partial x}\left(A(u)\frac{\partial u}{\partial x}\right) - h\frac{\partial^4 u}{\partial x^4} \Leftrightarrow u_{2t} = A'(u)u_x^2 - A(u)u_{2x} + hu_{4x}$$

The function ***(Au) (Au)*** describes the square of the pulse velocity in the membrane while the last term, proportional with the fourth-order derivative of the density variation, describes dispersion, with ***hh*** a positive parameter indicating dispersion's magnitude.

We will find traveling wave solutions of this equation considering a supplementary requirement, namely by imposing to have a polynomial expression: u'=V(u)u'=V(u). This is in fact the attached flow equation.

The similarity reduction based on Lie symmetry method will be applied to the Gierer-Meinhardt model for local excitation and global inhibition. It is expressed in the following form:

$$\frac{\partial u}{\partial t} = u^2 v - \mu u + \alpha\frac{\partial^2 u}{\partial x^2}$$
$$\frac{\partial v}{\partial t} = c - u^2 v - v + \beta\frac{\partial^2 v}{\partial x^2}$$

Here *u(x,t) u(x,t)* and *v(x,t) u(x,t) u(x,t)* represent the population densities of the activator and the inhibitor at time $t > 0t > 0$ and spatial location *xx* . When the Lie symmetries are computed, one finds that the system allows an invariant of the form $\xi = x \pm \lambda t. \xi = x \pm \lambda t.$ So, traveling wave solutions can be pointed out.

Keywords:
Generalized Boussinesq equation, Gierer-Meinhardt model, similarity reduction, traveling wave solutions.

*Corresponding author, e-mail: maria.alina2009@yahoo.com

Investigating interplay of intracellular regulation and population dynamics in a bacterial restriction-modification system

Stefan Graovac[1,2*], Andjela Rodić[1], Marko Djordjević[1]

[1] *Faculty of Biology, 2 Faculty of Physics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*

Abstract:
In-vivo dynamics of protein expression in bacterial cells, depends not only on intracellular regulation, but also on the relevant population dynamics, e.g. on the rates of cell and plasmid division that can change with time. In addition to complicating qualitative understanding of the underlying regulation, these population effects can also significantly increase dimensionality of the parameter inference problem, as they effectively couple the dynamics of the proteins that otherwise do not regulate each other. We consider this problem on a relatively simple case of bacterial restriction-modification systems, where we exploit one of the first available measurements of in-vivo expression of the restriction enzyme and the methyltransferase in these cells, which are under control of a specialized transcription factor (control protein, being the third gene in the system).

To model the system dynamics, we use a biophysical model of the system gene expression regulation that we previously developed, to which we now include complete cell population dynamics effects, due to both dividing cell and plasmid population. We resolve this problem iteratively. We find that including the dynamics effects, clearly improves the agreement of the model with the experimental data. Moreover, we find that neglecting such effects significantly distorts the predicted dynamics, and can lead to falsely identifying (or falsely neglecting) regulatory mechanisms that otherwise do not exist. Furthermore, we also systematically perturb the population dynamics, to test the effect of such changes on the protein synthesis dynamics. Consistently with previous results, such perturbations significantly change kinetics of the protein synthesis. However, and most importantly, an early peak in methyltransferase protein expression remains robust in these perturbations, which is clearly related with the main functional constraint to protect the host cell from cutting by the restriction enzyme. This clearly aligns with the idea of robustness of intracellular networks, where we now also propose that the regulatory interactions may be hardwired to preserve the main system dynamical constraints with respect to changes in the population dynamics – these changes might be expected in the natural environment, due to changes in the external conditions.

Keywords:
intracellular regulation, cell population dynamics, computational systems biology, biophysical modeling, restriction-modification systems

This work has been funded by MESTD, Republic of Serbia (OI173052).

*Corresponding author, e-mail: stefan.graovac@bio.bg.ac.rs

Predicting CRISPR/Cas associated small RNAs and their role in bacterial virulence

Jelena Guzina[1*], Wei-Hua Chen[3,4], Tamara Stankovic[2], Magdalena Djordjevic[5], Evgeny Zdobnov[3], Marko Djordjevic[1]

[1]*Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Studentski trg 3, Belgrade, Serbia*
[2]*Multidisciplinary PhD program in Biophysics, University of Belgrade, Studentski trg 1, Belgrade, Serbia*
[3]*Swiss Institute of Bioinformatics and Department of Genetic Medicine and Development, University of Geneva, Geneva, Switzerland;*
[4]*Current address: College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China*
[5]*Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia*

Abstract:
In addition to well-acknowledged CRISPR/Cas immune function, that enables clearance of foreign DNA from bacterial cells, the growing body of evidence suggests that the system can also influence the expression of endogenous genes, primarily related to bacterial virulence. This emerging non-canonical system functioning depends on the activity of small associated RNAs, encoded outside the CRISPR array (scaRNAs) – which act in complex with another small CRISPR/Cas-associated RNA species (tracrRNAs). The first step towards understanding non-canonical CRISPR/Cas activity is, therefore, detection of effector small RNAs, which is experimentally hardly feasible due to the scarcity of bacterial RNA-Seq data. In line with this, scaRNA was, up to now, identified in only one, relatively rare CRISPR/Cas subtype (IIB), while the prevalence of scaRNA:tracrRNA pairs throughout remaining Type II systems is currently inconclusive.

In this study we bioinformatically predicted small RNAs associated with Type II CRISPR/Cas systems of (mainly) pathogenic bacteria, by starting with subtype IIB, where scaRNA was initially found. We detected small RNAs directly from bacterial genome sequence, by jointly predicting transcription signals – transcription start sites (TSS) and terminators – and homology to the CRISPR array for the predicted transcriptional units. Particularly, we employed our previously developed approach for predicting bacterial TSS – note that accurate TSS detection is the main limiting factor in accurate small RNA predictions As another level of evidence, we explored the conservation of our predictions across related bacterial species, and also their match to the available RNA-Seq data.

In addition to recovering the existing experimental evidence, and identification of scaRNAs throughout Type IIB systems, we identified scaRNA:tracrRNA pairs in a number of IIA/IIC systems. The appearance of scaRNAs in IIA/IIC systems strongly co-occurs with the harboring bacterial strains being pathogenic, which suggests both the ubiquitous presence of scaRNA:tracrRNAs throughout Type II systems, and also the underlying connection of associated non-canonical system functions with bacterial virulence. Interestingly, our results point to a previously unrecognized mechanism of native small CRISPR/Cas-associated RNA functioning, which, however, closely resembles the mechanisms underlying CRISPR-based biotechnological applications.

Keywords:
CRISPR/Cas, small RNA, scaRNA, tracrRNA, bacterial virulence, non-canonical CRISPR/Cas functions

This work has been funded by MESTD, Republic of Serbia (project OI173052) and SNSF (SCOPES project IZ73Z0_152297).

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

Numerical simulation of stent deployment procedure in patient specific coronary artery

Velibor Isailovic[1,2*], Igor Saveljic[1,2], Themis Exarchos[3], Oberdan Parodi[4] and Nenad Filipovic[1,2]

[1] *Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia*
[2] *Bioengineering Research and Development Center, 34000 Kragujevac, Serbia*
[3] *Foundation of Research and Technology Hellas -Biomedical Research Institute, Ioannina, Greece, University of Ioannina, Ioannina, Greece*
[4] *Instituto di Fisiologia Clinica, Consiglio Nazionale delle Ricerche (IFC CNR), Pisa, Italy*

Abstract:
The cardiovascular diseases are the most common cause of death today. One of very serious health problems is growing of plaque in coronary arteries. The occurrence of plaque leads to narrowing of blood vessels and, as a consequence, reduced blood flow through the blood vessel. The solution for this patient's state is implantation of stent inside narrowed blood vessel. Using that kind of wired implants, it is possible to restore blood flow and significantly improve the patient's health. Today, huge efforts and financial resources are invested in the stent development. Experimental research is the backbone of research related to the development of new stents and stent materials. Also, numerical simulations play very important role in stent design optimization. Numerical methods, like finite element method can be used to appropriately simulate real stent behavior. Such technics allow to significantly reduce the number of experiments. Further, this reduces the time and cost of stent development. Beside mentioned advantages, such numerical methods allow to get a look inside during the process of stent implantation. This paper represents pretty simple approach that can be used for modeling of stent implantation including interaction between stent and blood vessel.

Keywords:
atherosclerosis, computer simulation, plaque propagation, wall shear stress

This work has been funded by EC HORIZON2020 689068 SMARTool and MESTD, Republic of Serbia (projects III41007 and OI174028).

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

Modelling Hospital Infection Spread in the Polish Regional Healthcare Network

Andrzej Jarynowski[1], Vitaly Belik[2]

[1] *Interdisciplinary Research Institute, Wroclaw, Poland*
[2] *Institute for Veterinary Epidemiology and Biostatistics, Freie Universität Berlin, Germany*

Abstract:
Background & Aims: The spread of hospital-associated infections has been studied through computer simulations on multiscale network of possible path of pathogen transmission between hospitals or within hospital. The goal of analysis was to provide advice on regional infection control strategy. Increasing spread of antibiotic resistant pathogens encourages us to prepare risk assessment for Malopolska Region (Lesser Poland).

Data & Inspirations: The nosocomial infections registry (incidence data) of Malopolska Region has been investigated for 18 chosen alert pathogens. Incidences vectors for each county were used to build a spatial network, where nodes are counties locations and edges represent epidemiological similarity (scalar product of incidence vectors). Patient movement patterns from Lower Saxony and Stockholm County have been also investigated. Pseudo gravity function of human mobility (taking into account populations of counties and distance between them) with epidemiological similarity (defined by similarity in pathogens incidence) has been modelled as a proxy for Polish patient referral structure. Network was constructed on two scales: regional (inside Malopolska) and single hospital (in one of Malopolska hospitals). Network of contacts between patients within single hospital was captured by set of questioners, CAD maps integration, functional paths annotation, local vison as well as by processing data from the register of patient admissions and discharges from each hospital unit (wards, clinics, etc.), and microbiological laboratory test results.

Methods: Malopolska database contains more than 10 000 infection events during hospitalization in 2015 from around 60 hospitals. Clustering as well as spatial correlations using agglomerative hierarchical structure detection techniques were performed. On the level of single hospital, the risk of acquiring infection by patients have been assessed, based on the patient's incoming connections (different centrality measures).

Results: Our analysis indicates that one border county `Oswiecim' does not belong to the rest of regional network. Moreover, the similarity in the incidence of viral pathogens (epidemic spread) seems to decay with distance much stronger than in bacterial pathogens (endemic spread). On the level of single hospital, the most accurate risk function has been selected – temporal adjusted Page Rank algorithm from transmission trees. However, much simpler algorithm (betweenness from contact network) is almost as good as a heavily parameter-depended outperforming algorithm.

Conclusions: Our analysis and modeling of a regional healthcare network could facilitate decision making in infection control from a system perspective.

Keywords:
epidemics, networks, hospital infection, modelling

*Corresponding author, e-mail: vitaly.belik@fu-berlin.de

Circulating miRNAs as Potential Liver-Related Biomarkers

Tim Kacprowski[1,2,3,*], Sabine Ameling[2,3], Georg Homuth[2], Maik Pietzner[3,4], Julia Mayerle[5,6], Jens-Peter Kühn[7], and Uwe Völker[2,3]

[1] *Chair of Experimental Bioinformatics, School of Life Sciences, Technical University Munich, Maximus-von-Imhof Forum 3, 85354 Freising, Germany*
[2] *Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany*
[3] *DZHK (German Centre for Cardiovascular Research), partner site Greifswald, Greifswald, Germany*
[4] *Institute of Clinical Chemistry and Laboratory Medicine, University Medicine, Greifswald, Greifswald, Germany*
[5] *Department of Medicine A, University Medicine Greifswald, Greifswald, Germany*
[6] *Medizinische Klinik und Poliklinik II, Klinikum der LMU München-Grosshadern, München, Germany*
[7] *Department of Radiology and Neuroradiology, University Medicine Greifswald, Greifswald, Germany*

Abstract:
MicroRNAs (miRNAs) are short RNAs of approximately 22nt length that have profound effects on gene expression. Through their regulatory role, miRNAs facilitate dynamic cellular responses and are involved in the development of diseases. Besides their intracellular localization, miRNAs can also be found in body fluids, e.g., plasma. Plasma circulating miRNAs (pcmiRNAs) show great promise as easily accessible biomarkers. They exhibit exceptional stability in plasma samples and are easily detectable by standard techniques such as RT-qPCR.

We determined pcmiRNA profiles in more than 700 individuals from the population-based Study of Health in Pomerania-TREND employing Serum/Plasma Focus microRNA PCR Panels (Exiqon) in a RT-qPCR approach. We established a workflow for association studies investigating relationships between pcmiRNAs and phenotypes. In particular, we investigated the putative relation between pcmiRNAs and liver fat quantified by magnetic resonance imaging.

In multiple linear regression models, we identified strong associations of pcmiRNAs with liver fat and liver enzyme activities in plasma. Furthermore, we employed cross-validated random forests for predicting fatty liver disease (liver fat content ≥ 5%). We considered base models relying only on liver enzyme activities or clinical scores as features. The inclusion of pcmiRNAs as additional features improved the prediction performance.

Our results indicate the involvement of specific pcmiRNAs in biological processes correlated with the accumulation of liver fat. For the investigated individuals we also have other omics data available, such as genome-wide genotypes, whole-blood gene expression, plasma metabolome, plasma proteome, and gut mircobiome. Thus, we have the opportunity to investigate the interplay between liver fat accumulation and multi-scale omics. This will further deepen our understanding of the biology underlying liver fat accumulation and will eventually allow for precise systems medicine approaches to tackle fatty liver disease.

Keywords:
miRNAs, association studies, bioinformatics, machine learning, systems medicine

This work has been funded by the Federal Ministry of Education and Research (BMBF, grant nos.01ZZ9603, 01ZZ0103, 01ZZ0403 and 03ZIK012) and the Ministry of Cultural Affairs and the Social Ministry of the Federal State of Mecklenburg-West Pomerania.

*Corresponding author, e-mail: tim.kacprowski@wzw.tum.de

Developing an Integrated Genomic Profile for Cancer Patients Utilizing NGS Data

Kosvyra A.[1*], Maramis C.[1], Chouvarda I.[1]

[1] *Laboratory of Computing, Medical Informatics and Biomedical Imaging Technologies, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece*

Abstract:
Next Generation Sequencing (NGS) has led to a great volume and variability of sequencing data, resulting in heterogeneous data sources, such as Whole Exome Sequencing (WES) and RNA Sequencing (RNAseq). The challenge in this case is to discover innovative ways to explore and combine these heterogeneous data sources in order to create an integrated profile of a patient that will allow us to build the complete picture of a disease.

Chronic Lymphocytic Leukemia (CLL), a neoplastic blood disease with a strong genetic influence, was used as an example of this approach. The whole analysis was conducted using open data.

The integrated approach described in this paper links the various NGS sources with the patients and their clinical data. The resulting profile efficiently summarizes the large-scale datasets and correlates indicators arising from different data types. With the use of machine learning techniques and the association of these indicators with the clinical information, it has been possible to build efficient predictive models. This approach includes:

1. Exploration/Visualization, e.g. depiction of the different analyses' results for one individual and intra-person comparison by linking these results

2. Inter-group comparison, e.g. comparison between the results of different analyses for cases with different clinical characteristics such as groups of patients with mutated and unmutated IgVH & EGR2 gene

3. Deployment of predictive models, e.g. a disease outcome prognosis model that predicts whether the disease is going to be in indolent or aggressive condition, using a two-class random forest classification algorithm.

The final goal is to design a complete genomic profile of a cancer patient. The profile includes visualization of the results of WES and RNAseq analysis (specific variants and significantly expressed genes, respectively), integration/comparison of these results and a prediction regarding the outcome of the disease or the response to treatment .

Concluding, this integrated profile can contribute to the medical research providing new possibilities in personalized medicine and prognostic views. This innovative/exploratory data-driven approach attempts to make use of the big genomic data by summarizing and presenting them in a way that renders them easily usable and interpretable by health professionals.

Keywords:
bioinformatics, sequencing analysis, data mining

*Corresponding author, e-mail: aekosvyra@auth.gr

Establishing Benchmark Criteria for Single Chromosome Bacterial Genome Assembly

Timothy Krause[1*] and Joshua R. Herr[1,2,3]

[1]School of Biological Sciences, University of Nebraska, 1901 Vine St, Lincoln, USA
[2]Department of Plant Pathology, University of Nebraska, 406 Plant Science Hall, Lincoln, USA
[3]Center for Plant Science Innovation, University of Nebraska, 1901 Vine St, Lincoln, USA

Abstract:
Genome sequencing has become commonplace in the two decades since the first bacterial genome was completed. Despite the sequencing of tens of thousands of bacterial strains, adequate recommendations for the amount and types of sequencing data needed to optimize the recovery of single chromosomes from bacterial sequencing projects still do not exist. Confounding this is the ever-shifting output due to improvements in existing sequencing data generation (long-read technologies) and new assembly methods. Broad estimates for optimal coverage depth exist to recover complete bacterial genomes, but required sequencing depths across bacterial and archaeal phylogenies are needed for high-quality assembly and are not well understood. In addition, the capabilities of multiplexing (sequencing more than one sample simultaneously on one flow cell) for long-read sequencing platforms in order to recover complete bacterial chromosomes are poorly documented. In order to address these issues, we randomly sampled raw public sequencing reads at different increments of coverage across a phylogenetically diverse array of bacterial strains. The sequencing reads were then assembled and assembly quality was assessed. Here we demonstrate that assembly quality peaks at varying coverage depths in different bacterial lineages and that most sequencing efforts rely on short read data which provide incomplete genome assemblies. We also demonstrate that characteristics such as genome size, k-mer abundance, and GC content may affect optimal assembly parameters. Furthermore, we simulated long-read data to complement the short read data based on standard multiplexed read profiles and found that although limitations due to genome size and repeat complexity exist, long-read x8 multiplexed data are able to close many bacterial genomes without the need for additional short-read sequencing.

Keywords:
Genome Assembly, Sequencing, Comparative Genomics

This work has been funded by University of Nebraska-Lincoln.

*Corresponding author, e-mail: timothy.krause@huskers.unl.edu

Extracting survival-relevant subnetworks from multi-scale omics data with KeyPathwayMiner

Manuela Lautizi[1,2], Tim Kacprowski[1], Paolo Tieri[2,3], Jan Baumbach[1], Markus List [1*]

[1] *Chair of Experimental Bioinformatics, School of Life Sciences, Technical University Munich, Maximus-von-Imhof Forum 3, 85354 Freising, Germany*
[2] *Faculty of Information Engineering, Informatics and Statistics, La Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy*
[3] *Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni del Calcolo, Rome, Italy*

Abstract:
Biological interaction databases can be exploited by pathway-level enrichment methods for downstream analyses in biological and biomedical settings. Classical enrichment methods rely on predefined lists of pathways, biasing the search towards known pathways and risking to overlook unknown, yet important functional modules. To overcome this limitation, so-called de novo network enrichment approaches extract novel pathways from large, molecular interaction networks given molecular profiles of patients, e.g. gene expression, promoter methylation, etc.

Network enrichment of molecular profiling data is challenging due to noise and incompleteness of both the data themselves and the networks. KeyPathwayMiner (KPM) jointly considers multi-scale molecular profiles to extract subnetworks enriched for de-regulated genes, e.g. differentially expressed genes. KPM is available as a feature-rich, user-friendly Cytoscape app, standalone software, or web service for de novo network enrichment (http://www.keypathwayminer.compbio.sdu.dk/).

Clinical cancer research often focuses on patient survival times. Thus, we developed a new strategy to identify subnetworks most significantly associated with differences in survival. Our approach is based on the Network of Mutations Associated with Survival (NoMAS) algorithm that extracts subnetworks enriched in mutations. NoMAS exploits colour-coding to identify candidate subnetworks that are then evaluated with a log-rank test. We adapted NoMAS for multi-scale omics data by introducing a k-means clustering step to split patients into two groups using the candidate subnetworks molecular profile. Next, we apply a log-rank test to assess the significance of the difference in survival times between the two groups. Our overall goal is to find subnetworks significantly associated with survival time, thus creating multi-scale models that connect molecular changes, e.g. on the level of gene expression, to changes in the time-scale of patient survival. The identified subnetworks can be expected to represent important disease mechanisms, making them interesting candidates for further investigation. We thus expect that extending KPM to survival data will make *de novo* network enrichment considerably more attractive as a systems medicine approach.

Keywords:
network enrichment, survival analysis, colour-coding, systems medicine

*Corresponding author, e-mail: markus.list@wzw.tum.de

Network-constrained bi-clustering of patients and multi-scale omics data

Olga Lazareva[1,2], Simon J. Larsen[3], Paolo Tieri[2,4], Jan Baumbach[1,3] and Tim Kacprowski[1*]

[1]*Chair of Experimental Bioinformatics, School of Life Sciences, Technical University Munich, Maximus-von-Imhof Forum 3, 85354 Freising, Germany*
[2]*Faculty of Information Engineering, Informatics and Statistics, La Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy*
[3]*Computational Biology group, Department of Mathematics and Computer Science, University of Southern Denmark, DK-5230 Odense M, Denmark*
[4]*Consiglio Nazionale delle Ricerche, Istituto per le Applicazioni del Calcolo, Rome, Italy*

Abstract:
Recent advances in omics profiling technologies yield ever larger amounts of molecular data. Yet, the elucidation of the molecular basis of human diseases remains an unsolved challenge. The analysis of multi-scale omics data requires integrative bioinformatic tools capable of multi-modal computing and multi-scale modeling. Unsupervised learning approaches are frequently employed to identify biomolecules and pathways involved in specific diseases. However, classical clustering is hardly suitable to analyse, e.g., gene expression data conjointly with experimental conditions and molecular pathway information. Since we are interested in gene sets displaying a consistent behaviour across different conditions, both genes and samples have to be clustered simultaneously employing models respecting the heterogeneity of such multi-scale data. To this end, we aim for extending bi-clustering approaches by including information encoded in biological networks.

Methods
BiCluE (Sun et al. 2013) has been the first software package tackling the weighted bi-cluster editing problem. It provides an exact algorithm based on fixed-parameter tractability (FPT). The bi-cluster editing problem is formulated as a bi-partite graph connecting features and samples. We then transform this graph into a disjunct set of bi-cliques while minimizing the editing costs (e.g., number of edges to be added/removed). Even though BiCluE yields potent solutions in many scenarios such as novel genotype-phenotype associations in GWAS data, it does not consider intrinsic feature relationships, e.g., interactions between proteins or regulatory interactions between genes. Therefore, we propose an extension of the BiCluE algorithm by mapping molecular interaction networks onto the bi-partite graph such that we impose constraints that force bi-cliques to respect intrinsic feature relationships. This reduces the computational complexity from O(4k) to O(2k), with k being the cluster editing costs due to a drastic reduction of the search space. Additionally, this model straight-forwardly allows incorporation of multi-scale data depending on the integrated network.

Results and conclusions
We demonstrate the validity and efficiency of our extension to BiCluE on simulated data. In general, such network-constrained bi-clustering approaches do not only allow for more stable feature selection, they also lead to more coherent functional enrichment, improving interpretability with respect to systems biology and systems medicine while being straight-forwardly applicable to multi-scale omics data.

Keywords:
bi-clustering, fixed-parameter tractability algorithm, multi-scale data, unsupervised analysis

*Corresponding author, e-mail: tim.kacprowski@wzw.tum.dev

Genome-wide endogenous RNA networks highlight novel biomarkers in cancer

Markus List[1,2], Azim Dehghani Amirabad[2,3,4], Stephan Laggai[5], Sonja Kessler[5], Christina Schultheiß[5], Dennis Kostka[6], Alexandra K. Kiemer[5], Marcel H. Schulz[2,4]

[1] *School of Life Sciences, Technical University Munich, Maximus-von-Imhof Forum 3, 85354 Freising, Germany*
[2] *Max Planck Institute for Informatics, Saarland Informatics Campus, Campus E 1 4, 66123 Saarbrücken, Germany*
[3] *International Max Planck Research School for Computer Science, Saarland Informatics Campus, Campus E 1 4, Saarbrücken, Germany*
[4] *Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Campus E 1.7, 66123 Saarbrücken*
[5] *Department of Pharmacy, Pharmaceutical Biology, Saarland University, Campus C 2 3, 66123 Saarbrücken, Germany*
[6] *Departments of Developmental Biology and Computational & Systems Biology, University of Pittsburgh, 530 45th St., Pittsburgh 15201, U.S.A.*

Abstract:
The competing endogenous RNA (ceRNA) hypothesis motivates the existence of so-called sponges, i.e., genes that exert strong regulatory control via miRNA binding in a ceRNA interaction network. This poses a powerful mechanism for cancer to deregulate parts of the cellular transcriptional program through one or few key sponge genes. In particular non-coding RNAs may act as sponges in cancer to facilitate changes in transcriptional programs without the risk of lethal side effects caused by expressing a protein at abnormally high or low levels.

In spite of the importance of this phenomenon, we currently lack an efficient method for inferring sponge interactions on a genome-wide scale. Moreover, confounding factors such as large differences in sample numbers prevent comparisons across different cancer cohorts.

This motivated us to develop sparse partial correlation on gene expression (SPONGE), a method that is orders of magnitude faster than previous approaches and allows for the construction of genome-wide ceRNA interaction networks. SPONGE is the first method to compute empirical p-values efficiently based on a series of null models and can thus control for confounding factors that were underestimated in previous studies.

SPONGE enabled us to build the most comprehensive set of genome-wide ceRNA regulation models for 22 cancer types based on miRNA and gene expression data from The Cancer Genome Atlas. Our results reveal novel sponge genes, in particular non-coding genes, which are suitable as survival markers in different cancer types. We further conducted additional wet-lab and computational experiments to assess the quality of the inferred sponge interactions in liver cancer. Our results highlight the relevance of ceRNA network inference for clinical research, in particular considering the role of non-coding RNAs in cancer and the potential of targeting non-coding RNAs in personalized medicine.

Keywords:
competing endogenous RNAs, microRNA regulation, non-coding RNAs, biomarker discovery, partial correlation, cancer

*Corresponding author, e-mail: markus.list@wzw.tum.de

Bioinformatics pipeline used for Next Generation Sequencing analysis of predictive markers in hematological malignancies

Irena Marjanović[1*], Jelena Perić[1], Bojana Stanić[1], Nadja Pejanović[1], Bojana Lučić[1], Tatjana Kostić[1], Nataša Tošić[1], Maja Stojiljković[1], Teodora Karan Djurašević[1], Milena Todorović Balint[2], Nada Suvajdžić Vuković[2], Dragana Janić[3], Goran Rakocević[4], Miloš Popović[4], and Sonja Pavlović[1]

[1] *Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11010 Belgrade, Serbia*
[2] *Clinic for Hematology, Clinical Center of Serbia, Pasterova 2, 11000 Belgrade, Serbia*
[3] *University Children's Hospital, Tiršova 10, 11000 Belgrade, Serbia*
[4] *Seven Bridges Genomics, Omladinskih brigada 90, 11070 Belgrade, Serbia*

Abstract:
Background: Molecular genetic markers are of great importance in hematological malignancies for diagnosis, risk stratification, clonal evolution monitoring and therapeutic intervention. Next generation sequencing (NGS) technology enables detection of new molecular genetic markers, but at the end of library sequencing process, huge amount of data are created, representing a challenge in data analysis. In this research, we created specific bioinformatics pipeline - software programs utilized at a particular step of data analysis of predictive markers in primary lymphoma of central nervous system (PCNSL) and acute myeloid leukemia (AML).

Aims: The aim of data analysis is to create bioinformatics pipeline in order to convert the raw data into results that can be seen in a visualization software and that can be easily interpreted. Aim of this research was detection of new somatic variants in PCNSL and AML using specifically created bioinformatics pipeline in NGS data analysis

Methods: Targeted, amplicon based, NGS technology was used. FASTQ files produced upon library sequencing were processed in four stages: basic quality control and trimming, alignment and preprocessing, additional quality control, variant calling and filtration. We used Illumina software and "in-house" developed tools. Resulted data were aligned in relation to the GRCh37 reference genome.

Results: NGS and bioinformatics analysis resulted in detection of 920 variants in coding regions in PCNSL patients, out of which 37 were new. In leukemia 412 (207 child AML and 205 adult AML) variants in coding regions were detected, out of which 12 were new. The average coverage of high-quality sequences was 2981x per amplicon.

Conclusions: Considering the amount of the raw data after library sequencing, bioinformatics has become an indispensable component in any laboratory performing a clinical NGS test. Creating or selecting adequate pipeline is a key in successful bioinformatics data analysis. Our research confirmed that the number of somatic variants in lymphoma is larger comparing to leukemia. Using NGS methodology and bioinformatics pipeline, we discovered new variants in PCNSL and AML samples, which represent new molecular genetic markers in these hematologic malignancies.

Keywords:
bioinformatics, next generation sequencing, hematological malignancies

This work has been funded by MESTD, Republic of Serbia (project III41004) and by European Commission, EU-FP7-REGPOT-316088, 2013-2016.

*Corresponding author, e-mail: irenas24@gmail.com

Information extraction from tables: case studies on extracting demographic information from tables in clinical trial literature and drug-drug interactions from tables in drug labels

Nikola Milošević[1*], Rob Hernandey[2],Cassie Gregson[2] and Goran Nenadić[1]

[1] *School of Computer Science, The University of Manchester, Kilburn building, Oxford road M13 9PL, Manchester, United Kingdom*
[2] *AstraZeneca plc, Cambridge, United Kingdom*

Abstract:
Tables in biomedical literature present important information, such as demographic information about patients in clinical trials, adverse drug reactions, potential drug-drug interactions or results of the studies. However, tables are often ignored by text mining methods that enable easy access for the researchers to the relevant information from the vast pool of published literature. Tables present information in a structured manner, in which the visual structure of the table in combination with the text and values presented in cells, inform how the table is read and what are the relationships between the presented information. The visual structure and visual relationships between presented information present additional challenges for the text mining systems dealing with tables.

We present the methodology that consists of 7 steps – (1) table detection, (2) functional analysis, (3) structural analysis, (4) pragmatic analysis, (5) semantic tagging (6) cell selection, (7) syntactic analysis. The methodology presents a hybrid approach with some steps utilizing machine learning approach, while the others utilize rule-based approach.

We validate our methodology on two different case studies. The first case study was extracting three types of demographic information (general information such as age and gender; respiratory tests, such as FEV1 and PEF; and quality of life questionnaire scores, such as asthma quality of life questionnaire and St. George respiratory questionnaire) from clinical trial documents about asthma and COPD from PMC articles. The second case study was extracting potential drug-drug interactions from structured product labels reported on DailyMed.

The methodology performed with precision, recall, and F1-scores in ranges of 0.75-0.96, depending on the task complexity, the complexity of table, and how standardized is the manner of presenting cues for a given variable.

Keywords:
text mining, health informatics, table mining, information extraction, clinical trials, drug-drug interactions

This work has been funded by EPRSC and AstraZeneca.

*Corresponding author, e-mail: nikola.milosevic@manchester.ac.uk

Numerical approach for determination of virtual functional assessment index in coronary artheries

Aleksandar Milovanović[1], Igor Saveljic[1,2], Themis Exarchos[3], Oberdan Parodi[4] and Nenad Filipović[1,2]

[1]1Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia
[2]Bioengineering Research and Development Center BioIRC, Kragujevac, Serbia
[3]Foundation of Research and Technology Hellas -Biomedical Research Institute, Ioannina, Greece, University of Ioannina, Ioannina, Greece
[4]Instituto di Fisiologia Clinica, Consiglio Nazionale delle Ricerche (IFC CNR), Pisa, Italy

Abstract:
Atherosclerosis is a blood vessel disease characterized by a decrease in lumen of the blood vessel due to which the blood and oxygen flow to a certain organ is reduced. This inflammatory disorder results in plaque formation, lipid accumulation and disturbing blood flow in arteries.

Fractional flow reserve (FFR) is a method used for assessing anatomical and physiological significance of coronary stenosis during invasive endovascular intervention. This highly scientific and expensive procedure also helps doctors in making decisions on coronary revascularization.

Virtual functional assessment index (vFAI), based on computational fluid dynamics modeling and three-dimensional (3D) coronary artery reconstruction, is an effective alternative to invasive FFR. The main advantage of vFAI is that it delivers physiological assessment without unnecessary invasive procedures by determining the hemodynamic relevance of a given coronary stenosis. Another benefit from this kind of numerical analysis is that it takes only few minutes and also helps patients save healthcare costs.

Using the finite element method and three-dimensional geometry of the coronary arteries, the values of the pressure change are calculated and, based on these values, virtual functional assessment index is determined and the results are presented.

Keywords:
fractional flow reserve, virtual functional assessment index, coronary artheries, finite element method

This work has been funded by EC HORIZON2020 689068 SMARTool, and MESTD, Republic of Serbia (projects III41007 and OI174028).

*Corresponding author, e-mail: acakg85@hotmail.com

Arithmetical Regularities Inside the Standard Genetic Code as a Clue for the Investigation of Natural Biocomputing

Nataša Ž. Mišić

Research and Development Institute Lola Ltd, Kneza Višeslava 70a, 11030 Belgrade, Serbia

Abstract:
An abstract formalization of natural biocomputing and its technical implementation in the form of a nature-inspired (bio)computing are long-term and extremely challenging goals. Numerous approaches have shown that a basic dichotomy between natural and artificial (bio)computing lies in holism versus reductionism, respectively. Thus, natural biocomputing is irreducible to the dualistic concept of hardware/software due to a strong structure/function relationship residing at different biomolecular scales. This biological property, among others (self-organizing, self-maintaining, self-replicating, scalability, complexity, adaptivity, evolvability, robustness…), is formalized within a descriptive theory of the cell – *autopoiesis*, as a coupling of a purely relational property with a topological property. Moreover, the internal nature of an autopoietic system/computing is crucially determined by the relational and *self-referential* quantities, and it is closely related to that of *relational biology*.

From the perspective of autopoietic computing, numerous arithmetical regularities obtained for the nucleon numbers of genetic code constituents (the canonical amino acids and nucleobases), based on the relational and self-referential quantities, can be accepted as a relevant evidence for *relational biocomputing*. Motivated by the research that is related to circular codes in the genome and the genetic code, as well as by the existence of nucleon balances for the Gamow's division of the Standard Genetic Code (SGC), here we further investigated a nucleon distribution related to the mentioned SGC properties and the Rumer's canonical nucleobase order. It has been shown that it is possible to execute the bipartition of SGC so that this two disjoint sets in relation to the total set satisfy the same *self-similar ratios* for the next different levels of reduced SGC by termination codons: the number of codons and the average frequency of single nucleobase, the total nucleon numbers of amino acid sets, the total nucleon numbers of RNA and DNA codon sets and consequently the total aggregate nucleon numbers of amino acid set and cognate codon set. Beside the scalability inside SGC, the relationship of some of these self-referential quantities to the structural characteristics of the DNA and RNA molecules are shown.

The results are discussed in the context of the genetic code origin and evolution, as well as of the natural and nature-inspired biocomputing

Keywords:
biocomputing, genetic code, nucleon distribution, nucleic acids, self-referentiality, scalability

This work has been funded by MESTD, Republic of Serbia (projects TR 32040, TR 35023).

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

GARLIC: A Bioinformatic Toolkit for Etiologically Connecting Diseases and Cell Type-Specific Regulatory Maps

Miloš Nikolić[1*], Argyris Papantonis[1], and Alvaro Rada-Iglesias[1,2]

[1] Center for Molecular Medicine Cologne, Robert-Koch-Str. 21, 50931 Cologne, Germany
[2] The Cologne Cluster of Excellence in Cellular Stress Responses in Aging-associated Diseases, Joseph-Stelzmann-Str. 26 50931, Cologne, Germany

Abstract:
Both common and complex human disorders (CCHDs) often show polygenic and multi-factorial etiology. Therefore, a vast effort has been made for uncovering the genetic basis of these diseases through genome-wide association studies (GWAS).

These studies reveal that 70-90% of all single nucleotide polymorphisms (SNPs) associated with CCHDs (e.g., diabetes, obesity, stroke, cancer) do not occur within genes (i.e., the are non-coding). Additionally, recent findings show that genetic variants occurring within regulatory elements may alter their regulatory properties and lead to quantitative changes in gene expression. However, the discovery of SNPs involved in CCHDs and of the pathological mechanisms underlying them is far from straightforward.

Here, we present GARLIC (GWAS-based Prediction Toolkit for connecting Diseases and Cell Type-specific Regulatory Maps), a user-friendly, multi-purpose software with an associated database and online viewer that, using a set of cis-regulatory regions obtained through different experimental techniques (e.g. DNase-seq, ChIP-seq, ATAC-seq) that can etiologically connect human diseases with relevant cell types based on the overlap between disease-associated genetic variants and cis-regulatory maps. Additionally, GARLIC can be used to retrieve potential disease-causative genetic variants overlapping regulatory sequences of interest, thus satisfying important needs within the field of medical genetics.

Keywords:
complex diseases, bioinformatics, GWAS, SNPs, database

This work has been funded by CMMC.

*Corresponding author, e-mail: m.nikolic@uni-koeln.de

MALDI-TOF/TOF and diagnosis of bacterial UTIs

Damir Oros[1], Marina Čeprnja[2], Jurica Žučko[1], Ena Melvan[1], Mario Cindrić[3], Amela Hozić[3], Jasenka Škrlin-Šubić[2] and Antonio Starčević[1*]

[1] *Faculty of Food Technology and Biotechnology, Zagreb University, Pierottijeva 6, 10000 Zagreb, Croatia*
[2] *Clinical Department of Laboratory Diagnostics, University Hospital Dubrava, Avenija Gojka Šuška 6, 10000 Zagreb, Croatia*
[3] *Division of Molecular Medicine, Ruđer Bošković Institute, Planinska 1, 10000 Zagreb, Croatia*

Abstract:
Urinary tract infections (UTIs) are the most common form of bacterial infections in the community and in hospitals, so fast and reliable microbial pathogen identification in human urine samples is required for clinical microbiology laboratory, especially for the world with more prevalent antibiotic resistance. In our study we analyzed 20 outpatient human urine samples with microbiologically proved presence of more than 105 CFU/ml with Matrix Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF/TOF). As a reference for bacterial diversity selected urine samples were determined using 16S rRNA gene sequencing.

The direct approach, without step in which the microorganisms are isolated and grown, provided successful identification of bacteria at the genus and species levels, especially for monobacterial samples of *Citrobacter koseri, Escherichia coli, Klebsiella spp.,* and *Proteus spp*. The results of this study demonstrate that we could identify different pathogen with MALDI-TOF/TOF from fresh or frozen, not cultivated human urine samples. In order to address problems encountered with hypothetical proteins during bioinformatics analysis with ProteinPilot we have developed our in-house software based on natural language processing.

Keywords:
native urine sample, pathogen identification, proteomics, genomics, bioinformatics

This work has been funded by the Croatian Science Foundation (project IP-06-2016).

*Corresponding author, e-mail: astar@pbf.hr

ragp: An R toolbox for mining plant Hydroxyproline rich glycoproteins

Danijela Paunović[1], Milica Bogdanović[1], Slađana Todorović[1], Ana Simonović[1] and Milan Dragićević[1*]

[1]Institute for Biological Research "Siniša Stanković", University of Belgrade, Department of Plant Physiology, Bul. despota Stefana 142, 11000 Belgrade, Serbia

Abstract:
Plant Hydroxyproline rich glycoproteins (HGRPs) comprise a highly diverse family of cell wall macromolecules, involved in a wide array of physiological functions such as cell expansion, somatic embryogenesis, self-incompatibility, signaling and pathogen responses. Due to biased amino acid composition, abundant in disorder promoting residues, HRGPs are intrinsically disordered proteins. The lack of a stable structure lessens the sequence constraints imposed on these proteins and hampers efforts for homology based identification. Current mining approaches, based on identifying sequences with characteristic motifs and biased amino acid composition, are limited to prototypical sequences.

Herby we present ragp, an R package for HGRP mining with a pipeline which emphasizes finding chimeric and short HRGP's which is especially useful for identification of arabinogalactan proteins. The ragp pipeline exploits one of HGRP key features, the presence of hydroxyprolines which represent glycosylation sites. These sites are identified using a gradient boosting model trained on plant sequences with experimentally determined hydroxyprolines, based on the local (21-mer) sequence around the target prolines. The model was validated on a set of sequences which were not used during the model building, as well as by using several resampling approaches. Apart from prediction of proline hydroxylation main ragp features include efficient communication with web servers for prediction of N-terminal signal peptides and GPI modification sites, sequence annotation by querying hmmscan, GO enrichment based on predicted pfam domains, and the ability to classify prototypical HRGPs.

ragp represents the first implementation of a HRGP mining workflow in the R statistical language. It implements common strategies for finding and classifying HRGP sequences along with an optional step where proline hydroxylation is estimated which leads to increased sensitivity and specificity of the filtered sequences. Since R is one of the leading bioinformatics platforms, the filtered sequences can be further analyzed by many specialized packages using the same environment.

Keywords:
bioinformatics, data mining, gradient boosting, hydroxyproline rich glycoproteins, arabinogalactan proteins

This work has been funded by MESTD, Republic of Serbia (projects TR31019, OI173024).

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

Prediction of Human Phenotype Ontology Terms For Intrinsically Disordered Proteins

Vladimir Perovic[1], Radoslav Davidovic[1], Branislava Gemovic[1], Neven Sumonja[1], Nevena Veljkovic[1]

[1]Centre for Multidisciplinary Research and Engineering, Vinca Institute of Nuclear Sciences, University of Belgrade, Mike Petrovića-Alasa 12-14, 11001 Belgrade, Serbia

Abstract:
Human Phenotype Ontology (HPO) standardizes clinical feature descriptions, in a way that is consistent and computer-readable. It uses a hierarchical scheme to assign clinical features to organ systems and allows association of diseases and human genes into hierarchical classes. Efficient predictions of gene-HPO term relationships will speed up the discovery of candidate disease genes and affected signaling pathways. In the focus of our research were intrinsically disordered proteins (IDPs) that act as hubs in interaction networks and play key roles in signaling pathways.

Consequently, dysregulation of IDPs emerges as a critical element for noncommunicable diseases, including cancer. In order to predict terms associated with a particular IDP gene we developed a model that transfers list of HPO terms enriched among predicted IDP interactors to each particular gene. Interactors are predicted by IDPpi, a method that predicts protein-protein interactions of this specific protein class according to sequence information.

We demonstrated that our method performs comparably well to the PHENOSTRUCT which is a state of the art method that utilizes domain knowledge. On the other hand, a significant advantage of our method is that it is universal and computationally more efficient.

Keywords:
human phenotype ontology, intrinsically disordered proteins, protein-protein interactions, prediction

This work has been funded by MESTD, Republic of Serbia (projects ON173001, ON173049).

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

Application of Machine Learning Algorithms to Detect Coronary Artery Disease using Genomic Data

Milos Radovic[1,2*] and Nenad Filipovic[1,3]

[1] *Bioengineering Research and Development Center, 34000 Kragujevac, Serbia*
[2] *Seven Bridges Genomics Inc, 11070 Belgrade, Serbia*
[3] *Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia*

Abstract:
Coronary artery disease (CAD) is one of the leading causes of death in European countries. For this reason, any development that may lead to successful diagnosis of the disease and its severity is of great importance. In addition, development of such diagnostic software tool may significantly decrease the healthcare costs, thus achieving one of the major goals in most European countries. In this paper, we utilized gene expression profiles of patients from the SMARTool project to build a machine learning based model for detection of CAD. The proposed approach consists of a feature selection followed by a classifier that stratifies patients based on the presence of CAD. For evaluation purposes, the K-fold cross validation procedure was used, where, in each iteration, observations belonging to the left-out fold were used for testing purposes (test set), while the remaining observations were used for feature selection followed by classifier training (training set). In each iteration of the cross validation procedure, we optimized hyperparameters of the model by applying nested cross validation procedure on the training set. Results obtained by the cross validation procedure indicate that the proposed approach may be used as a reliable diagnostic tool for the given problem domain.

Keywords:
Machine learning, feature selection, coronary artery disease, gene expression

This work has been funded by EC HORIZON2020 689068 SMARTool, and MESTD, Republic of Serbia (projects III41007, OI174028).

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

Investigating the role of key features in CRISPR-Cas system regulation

Andjela Rodic[1*], Bojana Blagojevic[2], Magdalena Djordjevic[2], Konstantin Severinov[3], Marko Djordjevic[1]

[1] *Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*
[2] *Institute of Physics, University of Belgrade, Pregrevica 118, 11000 Belgrade, Serbia*
[3] *Skolkovo Institute of Science and Technology, Nobelya Ulitsa 3, 121205 Moscow, Russia*

Abstract:
Activation dynamics of a native Type I-E CRISPR-Cas system, whose main function is to defend a bacterial cell from foreign DNA, has never been observed experimentally in *E. coli* because this system is silent under standard growth conditions and the signaling path leading to its induction is underexplored. Generally, *cas* genes encoding Cas proteins are transcribed from one promoter, while CRISPR array, containing samples of DNA sequences from past infections, is transcribed from another promoter into a long pre-crRNA transcript. This transcript is cut by the Cas6e protein into a number of small crRNAs which guide other Cas proteins to complementary foreign DNA which is consequently destroyed. From the experimental research thus far the main features of CRISPR-Cas regulation can be inferred: *i*) transcription repression by highly cooperative binding of H-NS protein molecules, *ii*) displacing the repressors by transcriptional activators (LeuO), *iii*) autoregulation of the activator LeuO, and *iv*) fast non-specific degradation of pre-crRNAs by an unidentified endonuclease.

We previously dynamically modeled pre-crRNA processing and showed that fast non-specific pre-crRNA degradation significantly contributes to the fast system OFF-ON transition. To investigate the role of the observed high cooperativity in transcription regulation, we propose and theoretically analyze a synthetic circuit for CRISPR-Cas activation in which *cas* genes are put under transcriptional control of a restriction-modification (R-M) system exhibiting qualitatively similar regulatory features. Namely, in AhdI R-M system, cooperative binding of C proteins, which can be outcompeted by RNA polymerase, represses transcription of their own gene and of restriction endonuclease gene. Using this approach we found that CRISPR-Cas regulation is optimized to achieve fast switching on and generating very large amounts of crRNAs, while fast pre-crRNA degradation additionally provides a delay in crRNA generation – which may be related with primed adaptation, i.e. ultimately evading autoimmunity.

Fast OFF-ON transition, and a delay in restriction endonuclease expression, seem to be the unifying dynamic principles of R-M system establishment in new host cells, regardless of their regulation mechanism. Therefore, R-M and CRISPR-Cas systems expression dynamics are probably constrained by the same task: to express toxic, but potentially useful molecules inside a cell.

Keywords:
CRISPR-Cas, restriction-modification system, gene expression dynamics, modeling gene regulation, synthetic gene circuit

This work has been funded by MESTD, Republic of Serbia (project OI173052) and SNSF (SCOPES project IZ73Z0_152297).

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

Application of multiscale smeared finite element model for modelling of mass transport in capillary systems and biological tissue

Vladimir Simic[1], Miljan Milosevic[1,2], Bogdan Milicevic[1] and Milos Kojic[1,3,4*]

[1] *Bioengineering Research and Development Center BioIRC Kragujevac, Prvoslava Stojanovica 6, 3400 Kragujevac, Serbia*
[2] *Belgrade Metropolitan University, Tadeusa Koscuska 63, 11000 Belgrade, Serbia*
[3] *Houston Methodist Research Institute, The Department of Nanomedicine, 6670 Bertner Ave., R7-117, Houston, TX 77030, United States*
[4] *Serbian Academy of Sciences and Arts, Knez Mihailova 35, 11000 Belgrade, Serbia*

Abstract:
Mass transport in living organisms is occurring from blood vessels to tissues (supplying oxygen, nutrients, etc.), and in the reverse direction (transporting waste products of cells). This process has been investigated in last decades by the use of computational methods. Due to functional complexity and heterogeneity of capillary systems, there is a need for simplified and robust computational models that address mass transport in capillary–tissue systems.

Recently introduced smeared modeling concept and composite smeared finite element (CSFE) offers accurate predictions for gradient-driven mass transport within tissue. In smeared modeling concept, transport from capillary system is smeared to continuous mass sources within tissue, where the basis for modeling transport through capillary walls is relation between capillary surface area and volumetric fraction. Also, the CSFE relies on the transformation of the one-dimensional (1D) constitutive relations into the continuum form expressed by Darcy's and diffusion tensors. Additionally, a field of correction function for diffusivity through the capillary walls is incorporated in smeared models, in order to have the same mass accumulation in tissue as in case of true 3D models. The parameters of the numerically determined correction function are: ratio of thickness and diameter of capillary wall, ratio of diffusion coefficient in capillary wall and surrounding tissue; and volume fraction of capillaries within tissue domain.

Extension of our smeared model includes the lymphatic system and transport within cells. The lymphatic vessels are treated in a way analogous to the capillaries, by introducing the corresponding Darcy and diffusion tensors. Organelles within cells are introduced as separate domains connected by their membranes to cell cytosol. The usual-measurable material properties of the continuum biological environments and biological barriers are used.

The smeared concept is implemented into our implicit-iterative FE scheme and into FE package PAK. Examples illustrate accuracy of the CSFE element, demonstrating robustness of the introduced methodology and its applicability to real physiological conditions

Keywords:
diffusion, convection, composite smeared finite element, lymphatic system, cell interior, correction function.

This work has been funded by MESTD, Republic of Serbia (projects III41007 and OI174028).

*Corresponding author, e-mail: mkojic42@gmail.com

Independent Multifactorial Association algorithm to assess genetic and neuroimaging features associated with neurodevelopmental domains

N. Vilor-Tejedor[1*], MA. Ikram[2], J. Sunyer[3], HH Adams[2], JR. Gonzalez[3]

[1]Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Barcelona, Spain; Universitat Pompeu Fabra (UPF), Barcelona, Spain; Barcelonabeta Brain Research Center, Pasqual Maragall Foundation, Barcelona, Spain.
[2]Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands; Department of Radiology, Erasmus MC, Rotterdam, the Netherlands
[3]Research Institute for Global Health (ISGlobal), Barcelona, Spain; Universitat Pompeu Fabra, (UPF), Barcelona, Spain; CIBER Epidemiologa y Salud Pública, Barcelona, Spain.

Abstract:
Introduction. Multivariate multiscale methods have the potential to better capture complex relationships that may exist between different biological levels. Multiple Factor Analysis (MFA) is one of the most popular methods to obtain factor scores and measures of discrepancy between data sets. However, singular value decomposition in MFA is based on PCA, which is adequate only if the data is normally distributed, linear or stationary. Also, including strongly correlated variables can overemphasize the contribution of the estimated components. **Methods**. In this work we introduced a novel method referred as Independent Multifactorial Analysis (ICA-MFA) to derive relevant features from multiscale data. This method is an extended implementation of MFA, where the component value decomposition is based on Independent Component Analysis. In addition, ICA-MFA incorporates a predictive step based on an Independent Component Regression. **Simulation study**. We evaluated and compared the performance of ICA-MFA with both, the MFA method and traditional univariate analyses, in a simulation study. We showed how ICA-MFA explained up to 10-fold more variance than MFA and univariate methods. **Application**. We applied and compared the impact of the novel algorithm in a study of 4,057 individuals belonging to the population-based Rotterdam Study with available genetic and neuroimaging data, as well as information about cognitive functioning. Specifically, we used ICA-MFA to detect relevant genetic features related to structural brain regions, which in turn were involved in the mechanisms of cognitive executive function. **Results**. The proposed strategy makes it possible to determine the degree to which the whole set of genetic and/or neuroimaging markers contribute to the variability of the symptomatology jointly, rather than individually. While univariate results and multimodal MFA combinations only explained a limited proportion of variance (less than 2%), our method increased the explained variance (10%) and allowed the identification of significant components that maximize the variance explained in the model. **Conclusions**. The potential application of the ICA-MFA algorithm, constitutes an important aspect of integrating multivariate multiscale data, specifically in the field of Neurogenetics.

Keywords:
Imaging Genetics, Neurogenetics, multiscale modelling, Rotterdam Study, simulation analysis

*Corresponding author, e-mail: natalia.vilortejedor@crg.eu

Bioinformatics analysis of correlation between protein function and intrinsic disorder

Goran Vinterhalter[1*], Jovana Kovačević[1], Gordana Pavlović-Lažetić[1]

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

Abstract:
The correlation of molecular function and protein intrinsic disorder is an important aspect of understanding the relationship between function, sequence and structure. This research was inspired by statistical correlation evaluation method described by Xie et al. [1] where authors studied the relationship between structure and function of proteins from Swiss-Prot database and where their function was described with Swiss-Prot function keywords. In this research, different set of proteins was used (CAFA3 challenge training dataset) and the function was described with molecular function terms of Gene Ontology. The results where compared with the original paper by first repeating the analysis with molecular function keywords and then by GO terms. We used PONDR VSL2b disorder predictor to label over 66000 CAFA3 proteins as putatively disordered or ordered while functional annotations were obtained from Swiss-Prot database. Out of 186 molecular function keywords with more than 20 annotated proteins, we found 53 to be highly order related and 44 highly disorder related. Under the same conditions, out of 1781 GO molecular function terms, we found 699 to be highly order related and 616 highly disorder related. GO terms results are presented as interactive graphs displaying complex hierarchical structure of Gene Ontology. Comparison of two functional annotations, GO and Keywords, showed consistent results when mapping was adequate, but comparison between original and new results revealed prevalence of binding related functions in new results (disorder related) even though they were not present in the original results. Because of the small number of mappings between keywords and GO terms, we propose a new method for deriving the missing mapping with the highest likelihood by measuring similarity (Jaccard index) between sets of protein annotated by different functions. We also showed that correlation between predicted disorder and sequence length can be effectively approximated by randomly generated proteins.

Keywords:
protein function, intrinsic disorder, Jaccard index

References:
[1] Hongbo Xie, Slobodan Vucetic, Lilia M. Iakoucheva, Christopher J. Oldfield, A. Keith Dunker, Vladimir N. Uversky, and Zoran Obradovic. "*Functional Anthology of Intrinsic Disorder. 1. Biological Processes and Functions of Proteins with Long Disordered Regions*". In: Journal of Proteome Research 6.5 (2007), pp. 1882–1898.

*Corresponding author, e-mail: gvinterhalter@gmail.com, {jovana,gordana}@matf.bg.ac.rs

Higher-order genetic interactions in prostate cancer and benign prostatic hyperplasia

Neven Sumonja[1], Ana Vukovic[2], Nikola Nonkovic[2], Branislava Gemovic[1], Zorana Nikolic[2], Dusanka Savic-Pavicevic[2], Nevena Veljkovic[1], Goran Brajuskovic[2*], Vladimir Perovic[1*]

[1]*Centre for Multidisciplinary Research and Engineering, Vinca Institute of Nuclear Sciences, University of Belgrade, Belgrade, Serbia*
[2] *Faculty of Biology, University of Belgrade, Belgrade, Serbia*

Abstract:
The aim of this study was identifying high-order epistatic effects in population-based study of prostate cancer (PC) and benign prostatic hyperplasia (BPH) patients in Serbia. The study used a previously published data set consisting of 21 candidate single-nucleotide polymorphisms (SNPs) in regulatory and noncoding regions of the genome, in miRNA genes and in RNA-induced silenceing complex genes. It involved 355 PC cases, 361 BPH cases and 360 healthy controls from Serbia.

In this work we applied empirical fuzzy MDR (EF-MDR) which overcomes limitations of multifactor dimensionality reduction (MDR), a method that has been widely used in detecting gene-gene interactions. Fuzzy MDR introduces uncertainty to simple binary high vs. low risk classification that is applied by MDR. In addition to that, EF MDR modification significantly reduces the execution time and simplifies parameter tuning.

In the analyzed datasets we detected a 4-order epistatic interaction connected with increased risk of BPH and different genetic interaction that is specific for PC sub-population. Further ontology searches and literature examination suggested unique pleiotropic relationships that underlay prostate cancer and benign prostatic hyperplasia in patient populations in Serbia.

Keywords:

*Corresponding author, e-mail: vladaper@vinca.rs; brajuskovic@bio.bg.ac.rs

Characterization of cellular metabolic response in abiotic stress-induced growth in Arabidopsis thaliana utilizing data mining

Brijesh Singh Yadav[1*], Daniel Chamovitz[2], and Shiri Frelich[3]

[1] *Department of Bioengineering, University of Information Science & Technology St. Paul the Apostle, Partizansk bb., street , 6000 Ohrid, Republic of Macedonia*
[2] *Department of Molecular Biology and Ecology of Plants, Tel aviv University, Israel*
[3] *Agriculture Research Organization, Newe Ya'ar, Research teams-Newe Yaar, Israel*

Abstract:
Abiotic stress-tolerant crop varieties are a long term goal of many research programs, thus the interest in understanding the progression of stress responses. We reanalyzed the AtGen Express transcription dataset to go beyond gene-level characterization, and to contextualize the discrete information into (1) a process-level signature of stress-specific, time-specific, and tissue-specific responses and (2) identify patterns of response progression across a time axis. To gain a functional perspective, ~1000 pathways associated with the differentially expressed genes were characterized across all experiments. We find that the global response of pathways to stress is multi-dimensional and does not obviously cluster according to stress, time or tissue. The early response to abiotic stress typically involves induction of genes involved in transcription, hormone synthesis and signaling modules; a later response typically involves metabolism of amino acids and secondary metabolites. By linking specific primary and secondary response pathways we outline possible stress-associated routes of response progression. The contextualization of specific processes within stress – tissue – time perspective provides a simplified representation of cellular response while reducing the dimensions in gene-oriented response description. Such simplified representation allows finding stress-specific markers based on process-combinations pointing whether a stress-specific response was invoked as well as provide a reference point for the conductance of comparative inter-plant study of stress response, bypassing the need in detailed orthologous mapping..

Keywords:
Abiotic stress, Cellular Pathways, Transcriptomic Response, Hormonal Signaling

*Corresponding author, e-mail: brijesh.yadav@uist.edu.mk

P7

Binding of metal ions and water molecules to nucleic bases

Jelena M. Andrić [1], Ivana M. Stanković [2], and Snežana D. Zaric [*3,4]

[1] *Innovation center, Department of Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia*
[2] *Institute of Chemistry, Technology and Metallurgy, University of Belgrade, Njegoševa 12, 11000 Belgrade, Serbia*
[3] *Department of Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Serbia*
[4] *Science Program, Texas A&M University at Qatar, Texas A&M Engineering Building, Education City, Doha, Qat*

Abstract:
Hydrogen bond interactions between nucleic bases and water have been analyzed by surveying Protein Data Bank (PDB) and quantum chemical calculations, for the noncoordinated and coordinated water molecule. Python home-made script was used for the PDB search and Gaussian09 program was used for the calculations of interaction energy at MP2/def2-QZVP level on model systems. The results from PDB survey and from the calculations are in agreement: hydrogen bonds are shorter and stronger when water is bonded to a metal ion.

The PDB search data show that metal ions prefer to bind nucleic acids through water than directly by coordination. Mg, Sr and Na are the most common metals to interact with nucleic bases through water molecule. It was shown that 19.2 % of metal-nucleic interactions are double hydrogen bonds and Mg and Sr are the most prone to form multiple hydrogen bonds with nucleic bases.

The calculated interaction energies for [Mg(H2O)6]2+ complex are in the range of (-12.94 to -49.96) kcal/mol, and for [Na(H2O)6]+ complex in the range of (-6.66 kcal/mol to -19.63) kcal/mol, while the interaction energies for noncoordinated water are (-4.63 to -8.93) kcal/mol. These calculated values for water-nucleic base hydrogen bonds are comparable to the strength of hydrogen bonds between nucleic bases (-5 to -47 kcal/mol) calculated previously, so these results may be relevant to understand the role of water molecules and metal ions in the process of replication and stabilization of nucleic acids and also toxicity.

Keywords:
nucleic base, metal, hydrogen bond, Protein Data Bank

This work has been funded by MESTD, Republic of Serbia (project 172065).

*Corresponding author, e-mail: szaric@chem.bg.ac.rs

P10

Neural Networks Implemented on Aorta with Abdominal Aneurism

Milos Anic[1,2,*] and Nenad Filipovic[1,2]

[1] *Faculty of Engineering, University of Kragujevac, FINK, Sestre Janjic 6, Kragujevac, Serbia*
[2] *Bioengineering Research and Development Center, BioIRC, Prvoslava Stojanovic 6, Kragujevac, Serbia*

Abstract:
Experimental procedures for determining the risk from abdominal aortic aneurysm (AAA) rupture are still considering usually only the diameter of aorta as a critical factor for rupture, while the module of elasticity, which differs from person to person, is not taken into account. Thus, it is more reliable to use maximum wall shear stress as a critical factor for rupture of AAA. This paper presents one of the possible ways to determinate the maximal stress using artificial neural networks (ANN) on the models of AAA, which are generated in software PAK-CAD. The aim of this paper is prediction of critical wall shear stress, and its location, which represent the indicators of the place where aorta may rupture.

Parameters of AAA which were taken as inputs of ANN with 5 different geometrical parameters used to create 100 models of AAA in software PAK-CAD. This study shows 4 different neural networks which were used to predict 4 different outputs (maximum wall shear stress and location or coordinates (x, y and z) respectively).

The results have been shown that mean absolute error of detecting maximum wall shear stress is 0.0011, while mean absolute errors of detecting coordinates are 0.603, 0.625 and 0.168 for x, y, and z coordinate respectively.

Keywords:
AAA, ANN, detection, maximum wall shear stress, location

This work has been funded by MESTD, Republic of Serbia (projects III41007 and 174028) and by EC HORIZON2020 777119 InSilc.

*Corresponding author, e-mail: anicmilos95@gmail.com

P9

Molecular docking study on the interaction of human procalcitonin with 3-(1-(2-mercaptoethylamino) ethylidene)-chroman-2,4-dion

Edina Avdović[1*], Dejan Milenković[2], Jasmina M. Dimitrić Marković[3], Srećko R. Trifunović[1], and Zoran Marković[2,4]

[1] *University of Kragujevac, Faculty of Science, Department of Chemistry, Radoja Domanovića 12, 34000 Kragujevac, Republic of Serbia*
[2] *Bioengineering Research and Development Center, Prvoslava Stojanovića 6, 34000 Kragujevac, Republic of Serbia*
[3] *Faculty of Physical Chemistry, University of Belgrade, Studentski trg 12-16, 11000 Belgrade, Republic of Serbia*
[4] *Department of Chemical-Technological Sciences, State University of Novi Pazar, Vuka Karadžića bb, 36300 Novi Pazar, Republic of Serbia*

Abstract:
Coumarins or benzo-α-pirones are a very large and important class of heterocyclic compounds and have an important place in the chemistry of natural products. The structure of coumarin contains the fused pyrone and benzene rings along with the carbonyl group on the pyrone ring. Procalcitonin (~13 kDa) is a peptide consisting of 116 amino acids. Procalcitonin is enzymatically degraded into lower molecular weight peptides. The biological effect of this protein was proven in the study of Nyen et al. (1996) who showed that the elevated concentrations of procalcitonin can lead to sepsis which can be treated with the anti- procalcitonin antibodies.

Molecular docking analysis was carried out in order to identify the inhibition potency of the coumarin 3-(1-(2-mercaptoethylamino) ethylidene)-chroman-2,4-dion (ligand) against human protein, procalcitonin. The ligand was prepared for docking by minimizing their energy using B3LYP-D3BJ/6-311+G(d,p) level of theory. The inhibition activity was obtained for ten conformations of ligands inside the protein. This study showed that the molecular docking analysis is a very important tool in the analysis of the interactions of biologically important molecules and human procalcitonin, in this case.

The lowest values of Δ*G*bind and *Ki* are found for conformation 1. By analyzing the position of active amino acids, it can be concluded that ligand binds at the catalytic site of substrates by weak non-covalent interactions. The most prominent are H-bonds, alkyl-π and π-π interactions. Alanin and glutamin in positions 28 and 35 in the primary structure of procalcitonin chain have a predominant role as active inhibition sites, regardless of the conformation of investigated ligands. ALA28 forms one H-bond (2.09 Å length) with N-H group of the ligand, while GLN35 forms one H-bond (2.81 Å length) with S-H group of the ligand. TYR33 forms weak π-π interactions with benzene ring of the ligand. On the other hand, MET36 forms alkyl-π interactions with pyron ring of the ligand. These preliminary results suggest that the investigated compound might exhibit inhibitory activity against procalcitonin.

Keywords:
coumarin, molecular docking, procalcitonin, hydrogen bond, π-π interactions

This work has been funded by MESTD, Republic of Serbia (projects III41007, OI174028).

*Corresponding author, e-mail: edina.avdovic@pmf.kg.ac.rs

P9

Unobtrusive Human Activity Recognition

Milan Bjelica*, Nadica Miljković, and Mirjana Simić-Pejović

Faculty of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia

Abstract:
Human activity sensing and recognition is a multidisciplinary task involving (biomedical) data acquisition, big data analysis, machine learning, data mining, and classification. The inferred information can be used in security applications (e.g. surveillance, access/intrusion control), smart homes (energy efficiency), entertainment (augmented reality), remote healthcare (health monitoring, fall detection) and so on.

Traditional approaches to this area firstly relied on sensors/tags attached to human body. Being intrusive by nature, they not only asked for user cooperation – thus making them unsuitable for security applications – but also altered the way the subjects would usually behave without the sensors. Next generation systems mitigated these problems by using contactless, i.e. device-free acquisition of visual and/or acoustic signals; although these are easily obtained through surveillance systems, several new concerns arose. Not only are line-of-sight path and proper environmental conditions like lighting or noise required for system operation, but the risk of leaking human privacy brings significant non-technical concern. The latest generation of human activity recognition systems uses radio-based techniques that operate under non-line-of-sight conditions. Especially, the systems based on channel state information in ubiquitous wireless local area networks, e.g. those conforming to the IEEE 802.11n standard, draw much of the research attention. These systems use the existing home wireless networks and commercially available (off-the-shelf) devices – access points and networked terminals like mobile phones, tablets, laptops, etc. which are located in distinct points in space. Human activities bring small, but measurable variations into radio channel transfer function, which can be processed to infer information on such fine-grained movements like gestures, keystrokes, lip language, or breathing.

We implemented such indoor activity recognition system, relying upon a standard wireless network interface card and modified firmware. During the testing, an existing wireless network in an apartment was used – the router was placed in one room, and the receiving laptop in another. The results show that the human activity – walking, sitting, lying – causes observable and predictable changes in both the signal-to-noise ratio and phase data.

Keywords:
activity recognition, data mining, wireless networks

This work has been funded by MESTD, Republic of Serbia (project TR 32028).

*Corresponding author, e-mail: {milan, nadica.miljkovic, mira}@etf.bg.ac.rs

P1

Defining dynamical property observables which ensure efficient restriction-modification systems establishment in bacterial host

Bojana Blagojevic[1], Magdalena Djordjevic[1], and Marko Djordjevic[2*]

[1] *Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia*
[2] *Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*

Abstract:
Restriction-modification (R-M) systems are rudimentary bacterial immune systems, whose essential ingredients are restriction enzyme (R), which cuts specific unmethylated DNA sequences, and methyltransferase (M), which protects the same DNA sequences. Therefore, expression of R-M system has to be tightly regulated, to ensure adequate establishment in naïve bacterial host. R-M systems are organized in different architectures (convergent or divergent) and are characterized by different features, i.e. dissociation constants of dimerization, translation initiation rates, binding cooperativities. We proposed that R-M systems should display the following dynamical properties: i) time delayed expression of R with respect to M, ii) rapid consecutive expression of R, iii) increased stability of R steady-state levels. The open question is how different R-M system architectures and features relate to the constraints on dynamical response, especially since experimental approach to the problem is very demanding.

To that end, we computationally addressed the above question, by applying our biophysical model of gene expression regulation in R-M systems on both: convergent and divergent representatives of the subset containing regulators called 'C proteins', and obtained that both systems display the same three dynamical properties. Additionally, we *in-silico* perturbed characteristic systems features, by either abolishing the existing or by introducing new ones, which adversely affected dynamical properties in both systems.

Our results imply that different R-M architectures and features may be explained by few simple dynamical properties of the system, providing a unifying framework for understanding these seemingly different systems. Furthermore, we developed the first theoretical model for expression dynamics of a divergent R-M system, which instead of possessing any of the convergent system features, has overlapping promoters as the main feature. We also provided predictions for perturbed R-M systems dynamics, which remain to be experimentally tested in the future.

Keywords:
restriction-modification, bacterial immune systems, transcription regulation, biophysical modeling

This work has been funded by the Swiss National Science Foundation under SCOPES project number: IZ73Z0_152297, and MESTD, Republic of Serbia (project 173052 and 171004).

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

P10

Phylogenetic analysis of putative Balkan specific mtDNA lineages

Slobodan Davidović[1*], Boris Malyarchuk[2], Jelena M. Aleksić[1], Miroslava Derenko[2], Andrey Litvinov[2], Milena Stevanović[1,3,4] and Nataša Kovačević-Grujičić[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11010 Belgrade, Serbia
[2] Institute of Biological Problems of the North, Russian Academy of Sciences, 18 Portovaya Street, 685000 Magadan, Russia
[3] Faculty of Biology, University of Belgrade, Studentski Trg 16, 11000 Belgrade, Serbia
[4] Serbian Academy of Sciences and Arts, Knez Mihailova 35, 11000 Belgrade, Serbia

Abstract:
Available studies on human populations in Europe demonstrate that increased levels of genetic diversity are present in populations from the Balkan Peninsula. In this region, frequent admixture and differential introgression of various gene pools were rather common during the entire history of humans. Furthermore, it has been shown recently that complex demographic history of human populations from the Balkans was characterized also by microdifferentiation of particular mitochondrial DNA (mtDNA) lineages that are potentially specific for this region.

In order to provide new insights into the presence and abundance of putative Balkan's mtDNA lineages we completely sequenced 164 mitogenomes from the Serbian population which occupies the central part of the Balkan Peninsula. We analyzed them phylogenetically agains available complete mtDNAs of modern and ancient Western Eurasians. Coalescence ages of subclades were estimated using the mutation rate for the entire mtDNA molecule.

Detailed analysis of complete mtDNAs found in Serbian population allowed us to describe several new mtDNA lineages and to delineate those that are potentially specific for the Balkan Peninsula. Along with K1a13a1 and X2q1, which have been identified previously as putative Balkan's lineages, we detected additional lineages potentially specific to this region - H6a2b, H7j, K1a4l, T1a1l, U1a1c2, U4c1b1, U5b3j and W3b4. The sample sizes of H6a2b, U1a1c2, U4c1b1, U5b3j and K1a4l, however, were too small, and thus, further analyses, with increased sample sizes, are required to support their Balkan's origin. Most the detected mtDNA lineages occur exclusively in south-Slavic populations of the Balkan Peninsula with their coalescence age estimates ranging from 0.9 to 3.9 kya. This suggests that some of these lineages most likely emerged after the settlement of Slavs in the Balkan Peninsula while the older ones probably evolved in indigenous populations which were present in this region prior to the Slavic settlement.

The obtained results further support the view that microdifferentiation of mtDNA lineages was present in the Balkans over time, and that further studies, with increased sample of complete mitogenomes from this region, are required to shed more light on evolutionary processes that shaped contemporary genetic landscape of Balkan's populations.

Keywords:
mtDNA, phylogeny, sequencing, Balkan Peninsula, Serbian population

This work has been funded by MESTD, Republic of Serbia (project 47025) and Russian

Foundation for Basic Research (grant 16-34-00014).

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

P5

Parametric Modelling and Computational Examination of Bicuspid Aortic Valve

Smiljana Djorovic[1,2*], Aleksandar Milosavljevic[3,4], Lazar Velicki[3,4] and Nenad Filipovic[1,2]

[1] *Faculty of Engineering, University of Kragujevac, Sestre Janjic 6, 34000 Kragujevac, Serbia*
[2] *Research and Development Center for Bioengineering BioIRC, Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia*
[3] *Institute of Cardiovascular Diseases, Institutski put 4, 21204 Sremska Kamenica, Serbia*
[4] *Faculty of Medicine, University of Novi Sad, Hajduk Veljkova 3, 21000 Novi Sad,Serbia*

Abstract:
Bicuspid Aortic Valve (BAV) is one of the most common congenital aortic valve malformations, where 33% of BAV patients may experience vascular complications. BAV is differentiated from an initially Tricuspid Aortic Valve (TAV). In BAV anatomy, fusion of two leaflets results in one larger leaflet which leads to the presence of two unequal-sized leaflets instead of three normal TAV leaflets. In clinical assessment, appearance and development of BAV complications (aortic stenosis, regurgitation, aneurysms of valvular sinuses and ascending aorta), due to alternated blood flow and malformed anatomy, have not been completely understood yet. For this reason, a computational approach can provide important information about the biomechanical characteristics of BAV and their influence on its behavior. In regard to that, the aim of this study was to computationally examine BAV biomechanics using structural analysis based on the Finite Element Method (FEM).

A parametric three-dimensional finite element model of the aortic root with BAV was created. The model's geometry, including the leaflets, raphe, commissures, sinuses and annulus, was based on the measurements from two-dimensional computed tomography images. The model was simplified by reducing the parameters to a smaller set which can define the curves and overall geometry. A dynamic structural FEM analysis, employing *in-house* PAK solver, was performed to simulate the cardiac cycle prescribing physiological time-dependent pressures to aortic root constituents, and considering average tissue properties.

BAV biomechanics was analyzed in terms of valve opening and closure in conjunction with computation of stresses acting on the aortic root components. In the systolic phase of the cardiac cycle, BAV was asymmetrically opened with an eccentric elliptic orifice. A more eccentric opening yields to a more severe functional stenosis. For the peak aortic pressure, the raphe region and commissures were under the notably higher Von Mises stresses than the surrounding areas, which may lead to further progress of BAV malformation.

The presented parametric approach and computational study can facilitate a methodical investigation of the effect of BAV pathology and mechanical loading on the major BAV parameters, which is pivotal for the optimization of surgical procedures aimed at restoring a normal aortic root function.

Keywords:
bicuspid aortic valve, parametric modelling, structural analysis, finite element method

This work has been funded by MESTD, Republic of Serbia (project III41007 and 174028) and EC HORIZON2020 689068 SMARTool.

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

P9

Parallelization of software for stent deployment inside artery

Tijana Djukic[1*], Igor Saveljic[1,2], and Nenad Filipovic[1,2]

[1] *Research and Development Center for Bioengineering, BioIRC, Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia*
[2] *Faculty of Engineering, University of Kragujevac, Sestre Janjic 6, 34000 Kragujevac, Serbia*

Abstract:
Stent implantation is one of the endovascular treatments of dilatations and stenosis within blood vessels. Stent is a structure made of special metal alloys; it has a tubular shape and is highly flexible. During the deployment process, stent also causes deformation of the arterial wall and ultimately alters the flow through the vessel.

Numerical simulation of stent deployment can be useful because it enables detailed analysis of the process. Also, using numerical simulations, parameters of the process can be additionally adapted, to ensure better and more appropriate placement and design of the stent. This paper presents stent deployment simulation software, first within rigid and then within deformable artery. The stent deformation is modelled using the simplex deformable model and the deformation of the arterial wall is modelled using the simple strain model. The interaction between stent and arterial wall is modelled using the transfer of forces that are interpolated using the Dirac delta function.

However, due to the large number of nodes in meshes representing the stent and artery, the simulation lasts very long. In order to perform real-time simulations, special parallelization techniques have to be applied during software implementation. Principles of GPU (Graphics Processing Units) programming are used to enable software to use the high performance capabilities of modern graphics cards on personal computers. To show the benefits of this approach and the parallelization speed-up, the execution time of sequential version (running on regular CPU) and parallel version (running using the GPU – graphics card) is compared, for both cases – rigid and deformable arterial wall. Simulations are performed for a deployment of stent within the artery. Geometrical data are patient-specific and obtained using imaging techniques.

The developed software provides information about stent deployment within the artery in real-time and thus can be used in clinical practice to improve endovascular treatments.

Keywords:
real-time simulation, deformable stent, blood vessel, parallel computing, speed-up

This work has been funded by EC HORIZON2020 689068 SMARTool, and MESTD, Republic of Serbia (projects III41007, OI174028

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

P7

Review of (effective) data collection methods for data-driven personalized medicine

Dragićević Jelena[1], Stanojević Nikola[2]

[1] *Serbian Neuroscience Society, Bulevar Despota Stefana 142, 11000 Belgrade, Serbia*
[2] *Optimised Healthcare, 47 High Street, Barnet, EN5 5UW, Herts, United Kingdom*

Abstract:
In recent years, medicine and biomedical research have lived through an unprecedented spike in new discovery. This has been allowed, in part, by more widespread use of information technology and large amounts of data. That, in turn, has given rise to a new direction of thinking in medicine, dubbed Personalized (precision) medicine. This concept is focused on optimizing all aspects of healthcare based on an individual patient's data. Despite progress in the biomedical field, the current health care outcomes are still inadequate for many conditions.

To be effective, the personalized medical approach requires large amounts of data centered around a single individual, stemming from multiple different sources. Biomedical data comprise one dimension of data necessary for provision of personalized medicine. Data related to the lifestyle, habits, choices, nutrition, etc. are also crucial to creating fully personalized protocols with the highest possible efficiency.

In this paper, the existing data collection methods in use in the medical and wellness industry have been compared and their cost-effectiveness and ease of use was estimated. Some of the existing self-reported data collection methods and their "user experience" aspects were also examined and evaluated. The relative value of different methods was approximated based on: quantity of data collected; granularity (resolution) of individual data points; cost per single data point; commercial availability; user-friendly interface and ease-of-use; relative impact of obtained data.

Commercially available at-home sample collection kits combined with online reporting have shown to be the most effective methods for obtaining biomedical data sets, while the accompanying phonotypical as well as self-reported assessment data seem to be most effectively captured using several standardized questionnaires coupled with user-friendly and engaging interfaces.

To conclude, a hypothesis, based on the results, of what the most cost-effective combination of data collection methods would be, is presented.

Keywords:
personalized medicine, data collection, home testing kits, cost-effectiveness

*Corresponding author, e-mail: jelena.d.dragicevic.90@gmail.com, Ns@optimisedhealthcare.com

P4

Discrete simulation of electrospinning jet's evolution

Ilektra Ferouka[1,3]*, Tijana Šušteršič[2,3,4], Marko Živanović[2,3] and Nenad Filipović[2,3,4]

[1]*Ecole Centrale de Nantes, 1 rue de la Noe, Nantes, France*
[2]*Faculty of Engineering, University of Kragujevac, Sestre Janjic 6, Kragujevac, Serbia 3Bioengineering Research and Development Center (BioIRC), Prvoslava Stojanovica 6, Kragujevac, Serbia*
[4]*Steinbeis Advanced Risk Technologies Institute doo Kragujevac, Kneza Milosa 25, Kragujevac, Serbia*

Abstract:
The most important characteristics of the electrospun fibers are their internal morphology and their diameter. They both depend on polymer's parameters (concentration, molar mass, conductivity), but also on the process parameters (temperature, humidity, distance between electrodes, shape of collector, etc). The most critical phase for the final fiber diameter is the whipping phase. The motivation for this research is therefore to simulate electrospinning and analyze the effects of some of the parameters without the necessity to perform the experiments each time.

The computational modeling was performed using Matlab. The calculations of the polymer's behavior between the electrodes were based on the discrete model of Reneker. Thus, the jet was modeled as a system of beads connected by viscoelastic elements. Measurements of fibers' diameters have been made on microscopy images of four experiments, with ImageJ software and the plugin DiameterJ. The study includes parameter modifications in order to examine jet shapes and fibers' diameter when optimal and non-optimal parameters are used. Electrospinning parameters used for these experiments were 21G needle type; voltage pairs 10kV;14kV;18kV;14kV applied on the nozzle and 0kV on the collector and 0.2ml/h; 0.2ml/h; 0.3ml/h; 0.3ml/h flow rates, for each of the experiments respectively. Material properties for 15wt% PCL solution were adopted from literature values, as they could not be determined experimentally at this point.

The simulation results in Matlab show good agreement with reports from literature. Initial results also show that the jet shape during electrospinning is an indicator for the fibers' diameter and thus, the resulting fibers' quality. Future studies would include more experiments and analysis of more images to confirm the initial results. The ultimate goal is to establish the exact influence of the above-mentioned parameters on the fibers' diameter. This would allow to set rough values of electrospinning parameters, which could be further fine-tuned, in order to obtain the best fibers without the need to perform many experiments. The beneficial effect of such simulations are time gain, but also reduced material consumption, maintenance costs etc.

Keywords:
electrospinning, mathematical modeling, Reneker model, parameters variation

This work has been funded by the European Project H2020 PANBioRA (grant no 760921-2) and MESTD, Republic of Serbia (projects III41007, OI174028)

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

P3

Function Annotation Algorithm Based on Sequence Spectral Features: Evaluation on Human Transcription Factors

Gemović Branislava, Davidović Radoslav, Šumonja Neven, Perović Vladimir, and Veljković Nevena*

Center for Multidisciplinary Research, Institute of Nuclear Sciences Vinča, University of Belgrade, Mike Petrovića Alasa 14-16, 11001 Belgrade, Serbia

Abstract:
The significant gap between sequence information and protein function knowledge sets a major challenge in the field of bioinformatics on automatically predicting protein functions from the sequence. Development of protein function models require a trusted set of known function annotations to learn from, such as those stored in the Gene Ontology (GO). But, even in the case of human proteins, which are extensively curated, there is a considerable number of proteins with missing annotations. This incomplete state of biological ontologies does not only affect development of prediction models, but it also undermines realistic estimation of model performances. In the latter case, the delay of the GO annotation process may lead to falsely disregarding some protein functions that are published in literature, but not present in GO, and labeling predictions of these functions as False Positives (FP). This study aimed to explore how realistic is the evaluation of function prediction model based on the amino acid coding and Fourier Transform of protein sequences by taking into account not only GO, but also experimental results published in biomedical literature.

We analyzed 10,672 human proteins and predicted their annotations in GO sub-ontology Biological Processes (BPO). We singled out a subset of proteins for which our algorithm correctly predicted the majority of annotations in BPO, with no or small number of false negatives (Recall≥0.9). From this subset, we further selected 14 transcription factors and validated their FP predictions.

There were 69 predicted leaf terms and 449 corresponding parent terms that were considered FP according to the GO ground truth for proteins in the dataset. Manual curation of available biomedical literature resulted in identifying evidence for 68% of leaf terms. If we couldn't find evidence for the leaf term, we furthered our search on its parent terms. This way we found evidence for 86% of terms in the GO slims generated using predicted terms for each protein.

This case study suggests that the performance of our model is underestimated when considering only GO, at least for those genes with high recall values

Keywords:
functional annotation, Fourier transform, transcription factors, Gene Ontology, literature mining

This work has been funded by MESTD, Republic of Serbia (project 173001).

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

P6

Conditional Random Fields based approach for classification of the reactants in some metabolic reactions

Milana Grbić[1*], Savka Janković[1], Dragan Matić[1] and Gordana Pavlović-Lažetić[2]

[1] *Faculty of Natural Science and Mathematics, Mladena Stojanovića 2, 78000 Banja Luka, Republic of Srpska, Bosnia and Herzegovina*
[2] *Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*

Abstract:
Prediction of role of metabolites in metabolic reactions may be of great importance for further understanding of metabolic pathway. Each metabolic pathway is a series of chemical reactions which involve reactants, products and various intermediates. There are two basic types of biochemical reactions that are characterized by their presence in either anabolic or catabolic pathways, i.e. pathways with the ability to synthesize molecules with the utilization of energy or to degrade metabolites with releasing energy. Adenosine triphosphate (ATP) appears in both types of pathways, either as a reactant (anabolic pathway) or as a product (catabolic pathway). In anabolic pathways, ATP can be degraded to adenosine diphosphate (ADP) or to adenosine monophosphate (AMP), as well as to different forms of phosphate groups (free or binded with metabolite in a phosphorylation process). In catabolic pathways, reverse processes of synthesizing ATP from its lower forms ADP and AMP occur.

In this work we focused on predicting the metabolite role in a particular reaction, related to its involvement in the energy transfer or in the phosphorylation process. Each reactant or product are classified into one of eight classes: donor/acceptor of one or two phosphate group, free phosphate group(s), phosphate (compound built by binding one or two phosphate groups(s) with metabolite), lower forms of ATP in anabolic pathways, ATP synthesized in catabolic pathways, phosphate group which binds in catabolic pathways and the class containing the rest of metabolites.

We used Conditional Random Fields (CRF) based approach for classification. CRF is a discriminative probabilistic method commonly used for structured prediction. This approach takes into account both the position of the metabolite in the reaction and information about the role of other metabolites from the same reaction. For modeling the CRF feature functions, three different templates were used and analyzed: (1) unigrams and bigrams, (2) unigrams, bigrams and trigrams and (3) bigrams and trigrams. The proposed method was tested on a set of metabolic reactions of the yeast Saccharomyces cerevisiae. The obtained results indicated high level of accuracy of the proposed method.

Keywords:
prediction, conditional random fields, metabolite reactions

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

P6

RepeatPlus - program for finding repeats in nucleic acids and proteins

Ana Jelovic[1*], Nenad Mitic[2], and Milos Beljanski[3]

[1] *Faculty of Transport and Traffic Engineering, University of Belgrade, 305 Vojvode Stepe, 11000 Belgrade, Serbia*
[2] *Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*
[3] *Institute of General and Physical Chemistry, Studentski trg 12, 11000 Belgrade, Serbia*

Abstract:
Finding repeats in nucleic acids and proteins is of great importance in molecular biology and a large number of efficient tools have been developed. As the number of repeats found can be very large, our goal was to provide a tool that can precisely find all repeated sequences and filter them according to input arguments and outputs the results in a convenient way.

RepeatPlus Program provides statistical filtering according to input sequence length and number of repeat occurrences. Filtering of the found repeated sequences is based on given probability threshold. It also provides a large number of other options such as filtering related to ambiguous letters in input sequence, mappings according to different physicochemical characteristics of amino acids, and motif mask filtering. Console output and file output (in a number of different formats) are well structured and can be easily utilized by another program. Alternatively, the program can insert found repeats to an ODBC-compliant database, thus enabling Data Mining and/or Big Data approaches.

Proposed filtering enables data reduction to set of statistically important (repeated) sequences which is extremely useful especially in the case of processing large number of repeated sequences with relatively small length, and provides a significant gain in performance and quality of results compared to outputting all the found sequences.

The RepeatPlus program is implemented in Python/C++ and is freely available for public use

Keywords:
repeat, nucleic acids, proteins

This work has been funded by MESTD, Republic of Serbia (project 174021).

*Corresponding author, e-mail: a.jelovic@sf.bg.ac.rs

P8

Subtle transcriptomic signals in circulation after myocardial infarction might indicate the ventricular remodeling outcome

Ivan Jovanović[1], Tamara Djurić[1], Maja Živković[1], Milica Dekleva[2], Nataša Marković Nikolić[2] and Aleksandra Stanković[1*]

[1] VINČA Institute of Nuclear Sciences, Laboratory for Radiobiology and Molecular Genetics, University of Belgrade, Mike Petrovića Alasa 12-14, 11001 Belgrade, Serbia
[2] Clinical Hospital Center Zvezdara, Department of Cardiology, Faculty of Medicine, University of Belgrade, Dimitrija Tucovića 161, 11000 Belgrade, Serbia

Abstract:
Prolonged duration of left ventricular remodeling (LVR) after myocardial infarction (MI) leads to progressive alteration in the structure, shape, size and function of the heart representing maladaptive LVR (MLVR) that precedes development of heart failure which is life threatening. The aim of this study was to bioinformaticaly investigate the mechanisms that potentially protect from MLVR by analyzing transcriptome in MI patients with/without MLVR six months after the MI.

Peripheral blood mononuclear cells of 21 patients who suffered MI (12 without MLVR/9 with MLVR) were sampled 6 months after the insult. MLVR was defined as progressive LV dilatation with LV diastolic volume increase (>20%) together with preserved or declined global LV ejection fraction at 6 months follow up. Transcriptome data were obtained by employing Illumina iScan microarray technology. Gene Set Enrichment Analysis (GSEA) was used to detect concordant differences in a priori defined gene sets (Gene Ontology biological processes – GO:BP) between patients without MLVR and with MLVR. We adopted default GSEA settings and the significance of an enrichment score was evaluated by a 1000 permutation test with respect to phenotypes.

GSEA analysis revealed regulation of anoikis as the top enriched process in phenotype without MLVR (ES = 0.57, NES = 1.83, GSEA nominal p-value < 0.001). Investigation of the leading-edge subset, core genes that accounts for the gene set's enrichment signal, indicated that integrin beta 1 (ITGB1) had the highest rank metric score in the gene-set. None of the enriched gene-sets were significant after FDR correction, defining the signal of this biological process as subtle.

Top regulated *ITGB1* gene is one of the major integrins that are responsible for cell-cell and cell-matrix interactions that acts in both cellular adhesion and signaling. Anoikis programmed cell death is induced by the loss of cell–matrix interactions, so *ITGB1* gene upregulation in patients without MLVR indicate that negative regulation of anoikis might be protective against MLVR. Possibility to identify subtle transcriptomic signals in circulation makes this approach applicable in the research of new therapeutics for the protection of post MI heart failure development induced by MLVR.

Keywords:
transcriptome, microarray, GSEA, myocardial infarction.

This work has been funded by MESTD, Republic of Serbia (project OI175085).

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

P5

Motor Imagery Classification using H2O Machine Learning Platform

Natalija Katić[1*], Nadica Miljković [1], and Aleksandra Marjanović[1]

[1] *University of Belgrade - School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11000 Belgrade, Serbia*

Abstract:
The aim of this study was to define and implement appropriate algorithm for classification of electroencephalography (EEG) signals recorded on the surface of the scalp when the subjects imagined left or right hand movement (termed *motor imagery*).

Data from open source bnci-horizon-2020 database was used. Electroencephalography signals were recorded from C3 and C4 scalp locations following international 10-20 system that correspond to left and right motor cortex locations, respectively. The preprocessing had included filtering in the frequency domain in order to remove the noise originating from power hum, eye blinks, and head movements. The signal was transformed using Common spatial pattern algorithm in order to maximize variances of EEG signals from both sides of scalp and its main characteristics had been defined as features. Features from training set were used as training dataset, while the performance of the classifier was evaluated by making predictions on two test sets. Classifier based on Linear discriminant analysis had been constructed and it had predicted whether the subject had imagined left or right hand movement with the classification error 0.040 and 0.100 on the first and the second test set, respectively.

In order to verify the obtained result, Naive Bayes (NB) classifier and Generalized Linear Model (GLM) were constructed using H2O Machine Learning Platform. Obtained errors for application of GLM and NB were the same: 0.000 and 0.038 on the first and the second training sets, respectively. Possible improvements in classification by constructing different type of classifiers are achieved using H2O Platform, as it has been shown that construction of NB and GLM using H2O platform improved classification performance compared to LDA. It is assumed that most of the errors occur due to the lack of precise estimate of motor imagery onset. Future study should be focused on protocol improvements.

Keywords:
EEG, Generalized Linear Model, H2O, Linear Discriminant Analysis, Naive Bayes

This work has been partially funded by MESTD, Republic of Serbia (projects 175016 and TR32038).

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

P8

Population pharmacogenomic aspect of glucocorticoids response in Serbian population

Nikola Kotur[1*], Biljana Stanković[1], Vladimir Gašić[1], Branka Zukić[1], Kristel Klaassen[1], Sanja Srzentić[1], Ksenija Vučićević[2], Sonja Pavlović[1]

[1]*Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia*
[2]*Department of Physiology, Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia*

Abstract:
Pharmacogenomics represents one of the pillars of personalized medicine. Genetic variants that represent biomarkers of drug efficacy and toxicity vary considerably in frequency between populations. As a consequence, side effects or therapeutic effects of certain drugs might represent serious issue in some populations, and not in others. Also, the cost-effectiveness of pharmacogenomic testing in one country depends on pharmacogenomic disposition of the population. Bearing in mind the importance of population pharmacogenomics, we aimed to optimize procedure for population-genetic data extraction obtained during analyses of clinical exomes (coding region of all clinically relevant genes – 4813 genes) and to apply this procedure to analyze potential pharmacogenomic markers of glucocorticoid response in Serbian population.

We selected group of 100 individuals of Serbian origin whose clinical exomes were analyzed using TruSight One Sequencing Panel (Illumina) on Miseq system. The variants in pharmacogenes relevant for glucocorticoid response were selected using PharmGKB database and articles available on PubMed (combination of keywords: glucocorticoids, prednisone, prednisolone, dexamethasone, pharmacogenetics, pharmacogenomics, drug response). Eighteen variants that were associated with glucocorticoid response located in NR3C1, ABCB1, GSTP1, ADRB2 and TBX21 genes were included.

Population pharmacogenomic data mining (PoPDaM) pipeline was developed using Python, R and Shell (Bash) scripts. The input for PoPDaM pipeline comprised of 100 VCF files obtained after pre-processing of NGS data by built-in Illumina software. The first step was to process all VCF files and create one dataframe which contained genotyped information of all covered variants of each subject. In the second step, previously selected pharmacogenomic variants of glucocorticoid response were used for filtering the data. Next, descriptive statistical methods were applied to extract information regarding the frequency of each glucocorticoid response relevant variant. Prepared data were matched with 1000 Human Genome and PharmGKB database for selected variants: information regarding frequency of variants in other populations and level of evidence for each pharmacogenomic marker were integrated into final report file.

Result of the applied PoPDam pipeline should be used as a guideline when taking into consideration pharmacoeconomic aspect of pharmacogenomic testing for glucocorticoid drugs and as a guideline for future pharmacogenomic studies in Serbian population.

Keywords:
population pharmacogenomics, data mining, glucocorticoid drugs

This work has been funded by MESTD, Republic of Serbia (project III41004).

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

P9

Risk prediction of bladder cancer progression from gene expression data

Tamara Krsmanović[1,2], Katica Janković[1,2], Dragana Mitić Potkrajac[1,2], Robert B. Russell[3], Gordana Apić[1]

[1] *Metisox, Lakinrose Pioneer House, Vision Park, Histon, CB24 9NL Cambridge, UK;*
[2] *CCNet Scientific, Branimira Cosica 2, 21000 Novi Sad, Serbia;*
[3] *Cell Networks, University of Heidelberg, BIOQUANT Building, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany*

Abstract:
Approximately 80% of patients diagnosed with urothelial bladder carcinoma (UBC) have tumors in the mucosa (Ta, carcinoma *in situ*) or submucosa (stage T1). These categories are grouped as nonmuscle invasive bladder cancer (NMIBC). Depending on the stage, grade, number of tumors and tumor size, prior recurrence rate and presence of concurrent carcinoma *in situ*, NMIBC patients experience recurrence at one year in up to 61% of cases. Up to 17% of all NMIBC will eventually progress to muscle-invasive bladder cancer (MIBC) with increased risk for developing metastatic disease. MIB and NMIB cancers imply different therapeutic strategies thus progressive patients deserve careful attention. In this study we applied *in silico* systems biology approach on gene expression profiles from progressive and non-progressive T1G3 bladder cancer patients to predict gene expression signature of tumor progression. From the whole genome analysis, 512 identified genes were significantly and differentially expressed between patients with progressive and non-progressive T1G3 bladder cancer ($p<0.05$, $FC>2$). For further analysis our expert knowledge system utilized a network-based approach that considers manually curated prior knowledge of genes, proteins and chemicals and their associations with each other and to thousands of pathologies. Our expert knowledge system reported involvement of these dysregulated genes in urinary system associated toxicities including urinary bladder carcinoma, kidney carcinoma and mesangial cell apoptosis. Network analysis of urinary bladder carcinoma endpoint revealed involvement of 202 dysregulated genes from gene expression studies. Centrality analysis selected and ranked 66 of these genes by the importance in the mechanism of bladder cancer progression. Review of annotated public data on top 30 ranked genes further identified 13 genes associated with high grade bladder carcinoma, invasiveness and poor prognosis, the highest ranking being FN1, CCNE1 and MMP9. Using our expert knowledge system we identified a gene signature that potentially differentiate progressive from nonprogressive T1 bladder cancer. Further validation would allow translation of these data to better bladder cancer treatment and outcome.

Keywords:
gene expression, bladder cancer, risk prediction

*Corresponding author, e-mail: dragana.mitic@metisox.com

P4

Improving Clustering Performance in Microbiome Studies

Tatjana Lončar-Turukalo[1*], Nina Maljković[1], Sofija Pantović[1], and Sanja Brdar[2]

[1] *Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia*
[2] *BioSense Institute, University of Novi Sad, Zorana Đinđića1, 21000 Novi Sad, Serbia*

Abstract:
Background and aims: Analysis of the human microbiome variations between different body habitats and their interaction with humans at cellular and genetic level is relevant for understanding of microbiome related diseases. Prevailing technique for taxonomic identification in microbial communities is 16S rRNA sequencing. Taxonomy relies on grouping species with certain similarity level into operational taxonomic units (OTUs). Between sample comparisons are based on different beta diversity measures. This study evaluates performance of 24 beta diversity measures, ensemble clustering and semi-supervised approaches with dimensionality reduction in clustering microbiome samples.

Methods: Data from "Moving pictures of the human microbiome" study include 1967 microbiome samples from oral, skin and gut sites of one male and female, sampled over 396 time points. 16S sequences preprocessing comprised: demultiplexing, removing primer, and quality filtering. Sequences were clustered into OTUs by taxonomy assigner-UCLUST with a similarity threshold of 97% against Greengenes reference database. OTUs biome table summarizes taxonomy of samples as observations counts per sample. Each of 24 beta-diversity measures was evaluated using respective distance matrix as input to spectral clustering with automatic determination of local scaling factor. Ensemble clustering approach over obtained individual partitions was assessed for performance improvement. For further enhancements semi-supervised kernel-learning algorithm was explored using selected beta diversity measures combined with dimensionality reduction t-Distributed Stochastic Neighbor Embedding. Clustering performance was evaluated using Adjusted Rand Index (ARI), and stability using 50 repeated runs of the clustering procedure with subsampling.

Results: Beta diversity measure performing the best spectral clustering of samples according to body habitat and gender were: abundance Jaccard, Hellinger and Kulczynski distance with ARI 0.59±0.03, 0.58±0.02, and 0.57±0.02, respectively. Ensemble approach combining all beta diversity measures slightly outperformed individual approaches with ARI 0.62±0.02. Both Kulczynski and Hellinger affinity matrix combined with dimensionality reduction to 3 features, and semi-supervised clustering achieved ARI 0.84±0.02.

Conclusions: Beta diversity measures perform better if imply normalizing a taxonomic unit's contribution with sample's total observation counts. The best results achieved with dimensionality reduction impose the need for exploiting sparsity of OTU matrices. Clustering showed perfect separability with respect to habitat, while gender-wise skin samples could not be perfectly resolved.

Keywords:
microbiome, 16S rRNA, clustering, semisupervised learning, bioinformatics

This work has been funded by MESTD, Republic of Serbia (projects TR32040, III44002, III 44006).

*Corresponding author, e-mail: turukalo@uns.ac.rs

P8

Analysis of Amino Acid Interactions Based on Geometric Distances

Maljković Mirjana M[1*], Mitić Nenad S[1], and Beljanski Miloš V[2]

[1]*Department of Computer Science, Faculty of Mathematics, University of Belgrade, P.O.B. 550, Studentski trg 16, 11001 Belgrade, Serbia*
[2]*Bio-lab, Institute of General and Physical Chemistry, P.O.B. 45, Studentski trg 12/V, 11001 Belgrade, Serbia*

Abstract:
It is well established that amino acid interactions in a protein chain influence its three-dimensional structure, therefore knowledge about their spatial distribution is of importance for the understanding of mechanisms and predictions of protein folding. Using atomic coordinates for proteins described in the Protein Data Bank (PDB) database, a database with three types of distances between every pair of amino acids in the same chain was made. Using Euclidean metric, the following distances were calculated: distances between Cα atoms of amino acids, minimal and maximal distances without taking into account hydrogen atoms. Since descriptive statistics with graphics analysis are the basis of quantitative analysis of data, descriptive statistics for distances of amino acid pairs with different properties were calculated and the results were presented in 20x20 tables. The distribution of calculated distances of each amino acid pair was shown by graphics. Using different criteria (a type of distances, protein structural classification (SCOP class), secondary structure types of amino acids in pairs, intra-secondary or inter-secondary structure pairs and a threshold of calculated distances) different groups of amino acid distances were selected for analysis. For each group, descriptive statistics and graphic representations of distance distributions were calculated. The obtained data can be accessed via the website. The obtained data can contribute to the identification and better understanding of previously not observed unique interactions between amino acids in order to better predict the protein structure and function as well as to design new artificial proteins.

Keywords:
geometric distances, PDB, DSSP, SCOP, descriptive statistics, distance graphics

This work has been funded by MESTD, Republic of Serbia (projects III44006 and 174021).

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

P3

Correlation of intrinsically disordered protein regions content with environmental characteristic in Archaea and Bacteria

Saša N. Malkov[1*], Nenad S. Mitić[1], Gordana Pavlović-Lažetić[1], Jovana J. Kovačević[1], Miloš V. Beljanski2, and Mirjana D. Pavlović[2]

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

Abstract:
Intrinsically disordered proteins (IDPs) and intrinsically disordered (protein) regions (IDRs) do not adopt a well-defined (ordered) structure under physiological conditions. They are involved in essential cell processes through two basic mechanisms: the entropic chain mechanism which is responsible for rapid fluctuations among many alternative conformations, and molecular recognition via short recognition elements that bind to other molecules. A number of Archaeal and Bacterial proteomes are analyzed for intrinsically disordered protein region (IDR) content and the correlations of IDR content and organisms genome/proteome characteristics (such as GC percent, genome size, proteome size, average protein length) with environmental characteristics (such as motility, oxygen utility, temperature growth, habitat condition). Obtained results on individual environmental characteristics are in accordance with previously published results. More detailed analyzes of pair-wise and triple-wise correlations of different environmental characteristics with proteome IDR content provides some new observations. While pair-wise correlations show that environmental characteristics have mostly orthogonal impact on IDR content, some combinations have a stronger impact. For example, bacteria with psychrophilic temperature range have a low level of IDR content in aerobic and unknown oxygen requirement, as well as if they live in multiple or unknown habitats. On the other side, bacteria with facultative oxygen requirements have higher level of IDR content in unknown temperature range. IDPs and IDRs possess a high adaptive potential and may contribute to the improved adaptability of the organism by accommodating adaptive changes within short time frames.

Keywords:
bioinformatics, protein disorder, environmental characteristics

This work has been funded by MESTD, Republic of Serbia (project 174021).

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

P5

Classification of proteins into COG categories based on n-gram patterns

Ulfeta Marovac[1], Nenad Mitić[2*], and Miloš Beljanski[3]

[1] *State University of Novi Pazar, Vuk Karadžić bb, 36300 Novi Pazar, Serbia*
[2] *Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*
[3] *Institute of General and Physical Chemistry, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*

Abstract:
Proteins are significant biological macromolecules of polymeric nature (polypeptides), which contain amino acids and are basic structural units of each cell. Their contents include 20+3 amino acids and, as a consequence, they are presented in biological databases as sequences formed from 23 different characters. One of the possible classifications of proteins by their function is related to their contents in a certain cluster of orthologous groups (COGs). This classification is based on the previous comparison of proteins by their similarities in their primary structures, which is most often a result of homology, i.e. their mutual (evolutionary) origin. Experimental determination of a biological sequence function is a long and an expensive process.

In this paper is proposed the model for classification of proteins in COG categories based on amino acid n-gram patterns. The set of data contains protein sequences of genomes from Prochlorales bacteria which are known to have been classified by COG categories. The presented method is used for separation of n-gram patterns characteristic for proteins of corresponding COG categories. These patterns depend only on the amino acids context and the length of n-gram. Thus the presented method significantly reduces the number of processed n-grams in comparison to previously used methods of n-gram analysis. Therefore more memory space is provided and less time for protein procession is necessary. The number of possible n-gram patterns related to possible n-gram is about 10n-3 times smaller. The advantage of the new method is in its avoidance of sequence-sequence comparison and in search for those patterns (n-grams, up to 10 long) in a protein which are characteristic of the corresponding COG category.

On the basis of the proposed method, the predictor for determination of the corresponding COG category for a new protein is implemented. During the test phase, the constructed predictor shown satisfactory results, with the maximal precision of 99% reached for some proteins. According to its properties and relatively simple construction, the proposed method can be applied in similar domains where the solution of a problem is based on n-gram sequence analysis.

Keywords:
mining sequential patterns, cluster of orthologous groups, amino acids, n-gram

This work has been funded by MESTD, Republic of Serbia (projects III-44007 and 174021).

*Corresponding author, e-mail: umarovac@np.ac.rs

P7

In silico identification of Transcription End Sites in Human Genome

Sudheer Menon[1*], Keiko Ozato[2], and Ninoslav Marina[1]

[1] *University Institute for Information Science and Technology (UIST), Partizanska bb, 6000 Ohrid, Macedonia*
[2] *National Institute of Child Health and Human development (NICHD), Bethesda, Maryland, USA*

Abstract:
Human genome store complex information. Human genome contains 3 billion bases involves directly or indirectly the synthesis of cell, tissue or organ. We do not have a clear information about all protein coding, non-coding transcripts and genome elements that regulates gene expression. The next generation sequence data need to be analysed by using various existing and new computational tools. We have used several existing Next Generation Sequence (NGS) tools and developed certain tools and methods to understand vast variety of sequence data. Transcription End Sites (TES) controls gene expression. We have used PERL scripts to extract the Transcription End Sites form human genome and matched with the known gene annotation. This reveals several novel genes that codes for protein. We found 228 novel Transcription End Sites that acts together with RNA polymerase in gene activation, the remaining 67 sites have start/stop conflicts and the 143 codes for protein with known function.

Keywords:
next generation sequencing, coding and non-coding transcripts, transcription end sites

This work has been funded by National Institutes of Health (NIH).

*Corresponding author, e-mail: sudheer.menon@uist.edu.mk

P8

Inactivation of free radical species with selected triazoles

Žiko Milanović, Jelena Đorović[2*], Zoran Marković[3], Ana Amić[4] and Svetlana Jeremić[3]

[1] *Faculty of Science, University of Kragujevac, Radoja Domanovića 12, 34000 Kragujevac, Serbia.*
[2] *Bioengineering Research and Development Center, Prvoslava Stojanovića 6,34000 Kragujevac, Serbia.*
[3] *State University of Novi Pazar, Department of Chemical-Technological Sciences, 36300 Novi Pazar, Republic of Serbia*
[4] *Department of Chemistry, Josip Juraj Strossmayer University of Osijek, Cara Hadrijana 8a, 31000 Osijek, Croatia*

Abstract:
Many diseases such as cancer, inflammation, hypertension, and cardiovascular disorders can occur as a consequence of free radicals' action in the human body, due to the environmental and modern lifestyle factors. Most of the natural and artificial antioxidants cannot pass the blood-brain barrier, so the molecules locally produced are gaining more attention these days. More than 20% of oxygen is used in the brain, and it is clear that the species present there are constantly exposed to the reactive oxygen species (ROS).

Various quantum-mechanical methods are used to quantify the antioxidant activity in numerous studies. The reactions between the molecules with antioxidative properties and free radicals can follow two different pathways: H-atom abstraction and radical adduct formation. In this study, the evaluation of possible inactivation of the three free radicals (hydroxyl, hydroperoxyl, and chloromethylperoxyl radical) with chosen 1,2,4-triazole-3-thiones is performed. The reactions of antioxidant mechanism between the triazoles and mentioned radicals are investigated thermodynamically. The thermodynamic parameters that describe H-atom abstraction mechanisms are calculated and analysed. The equilibrium geometries of all studied compounds, radical cations, radicals and anions, as well as all other species that participate in the reactions of investigated mechanisms, were calculated using B3LYP-D3 functional in conjunction with 6-311++G(d,p) basis set. All calculations are performed in water, as a polar solvent. The obtained results show that all of the three investigated radicals can be neutralized with the examined triazoles, the only question is which mechanistic pathway to follow. Analyzing the results obtained for hydroxyl radical, it is notable that it can be scavenged via HAT and SPLET mechanisms. The lower values are obtained for ΔHBDE, indicating HAT as preferred mechanism of antioxidant action. In the case of hydroperoxyl radical, achieved ΔHPA values are lower than the corresponding values obtained for HAT mechanism. This fact makes SPLET mechanistic pathway the most favorable for antioxidant action. Regarding chloromethylperoxyl radical, there is present competition between HAT and SPLET mechanisms. As far as SET-PT mechanism is concerned, it is not a possible reaction pathway for inactivation of neither of the investigated radicals.

Keywords:
free radical species, antioxidant action

This work has been funded by MESTD, Republic of Serbia (projects III41007 and 174028).

*Corresponding author, e-mail: milanovicziko463@gmail.com

P2

Muscle model with net of fibers used for modelling cell migration

Bogdan Milicevic [1], Miljan Milosevic[1,2], Vladimir Simic[1] and Milos Kojic[1,3,4*]

[1] *Bioengineering Research and Development Center BioIRC Kragujevac, Prvoslava Stojanovica 6, 3400 Kragujevac, Serbia*
[2] *Belgrade Metropolitan University, Tadeusa Koscuska 63, 11000 Belgrade, Serbia*
[3] *Houston Methodist Research Institute, The Department of Nanomedicine, 6670 Bertner Ave., R7-117, Houston, TX 77030, United States*
[4] *Serbian Academy of Sciences and Arts, Knez Mihailova 35, 11000 Belgrade, Serbia*

Abstract:
Actin filaments, usually in association with myosin, are responsible for many types of cell movements. The most obvious movement is muscle contraction. Myosin is a protein that converts chemical energy in the form of ATP to mechanical energy, thus generating force and movement. However, interactions of actin and myosin are also responsible for movements of non-muscle cells in processes like cell division, cell migration etc. A cell contains net of fibers with actin and myosin, making it capable of contraction.

In biomechanics, one of the most used muscle models is Hill's model. Hill's equation describes the ability of muscle tissue to perform contractions. The equation describes relationship between stress, and velocity and stretch, and is mostly used to describe behavior of skeletal muscles.

We here introduce Hill's muscle model with net of fibers, in order to describe contraction of cell during cell migration. Cell migration consists of three phases: protrusion, contraction and relaxation. Cell is modeled by 2D elements with muscle fibers in direction of cell movement. Protrusion phase is generated by prescribing normal strains in direction of motion, with lateral strain which is ½ of the longitudinal normal strain. Activation function for muscle fibers within 2D continuum is prescribed during contraction phase. Once a cell starts to migrate, it is periodically contracting, and it can move between other cells.

Introduced concept is implemented into our implicit-iterative FE scheme and into FE package PAKSF. It is assumed in our finite element model that the fiber directions are specified by two points at each fiber, hence natural coordinates, initial lengths of fibers and fiber directions can be calculated, used further during calculation. Material model for continuum can be used independently of the muscle model. For cells we used Maxwell's material model with imposed incompressibility condition. It is assumed that the muscle fraction is given, and the total stress is calculated by summation of stresses in continuum and fibers. In numerical examples we demonstrate how cells migrate, through a continuum taken as viscous fluid and among other deformable cells.

Keywords:
Hill's muscle model, cell migration, finite element analysis

This work has been funded by MESTD, Republic of Serbia (project III41007 and 174028).

*Corresponding author, e-mail: mkojic42@gmail.com

P2

Geometry Optimization of Nitinol Stent Design based on FEA Topology Optimisation

Dalibor Nikolic[1*], Dejan Krsmanovic[2], Nenad Filipovic[1,3]

[1] *Bioengineering Research and Development Center, 34000 Kragujevac, Serbia*
[2] *CardioMed Technology Consultants, Miami, FL, 33014, USA*
[3] *Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia*

Abstract:
One of the problems in stent implantation is a process called *in stent restenosis* (ISR). In the pre-stent era, the occurrence of restenosis ranged between 32-55% of all angioplasties, and in bare-metal stent (BMS) era this range dropped to 17-41%. Many factors have influence on this phenomenon. Some studies show that in stent restenosis, strut shape and thickness have significant impact, especially if the stent is implanted in the small arteries. For better stent geometry modeling, we suggested novel approach – topology optimization on the existing stent design. For evaluation of the topology optimization and mechanical performance of nitinol stents as well as comparison of differences between old design (non-optimized) and new optimized design the finite element method was used. The Z-shaped closed-cell self-expanding stent type for testing was used. This type of stent is most common in clinical practice. Simulation was performed assuming that the stent devices used for this research were made by laser cutting, from tube form, by application of expanding and crushing force. The behavior of two different stent models was analyzed: old Palmaz-Schatz design and optimized design obtained based on the results from topology optimization of the Palmaz-Schatz design.

The main reason for using Palmaz-Schatz stent geometry is that this is a very simple geometry with enormous potential for modification. Performed simulation on stent models showed that the new modern design has better clinical behavior due to lower contacting surface, higher radial resistive strength and much better superplastic behavior. Optimization process was based on two optimization rules: minimization of the model volume and retention (or increase) of the maximal strain of the basic model.

Keywords:
stent, restenosis, computer simulation, optimization, topology optimization

This work has been funded by EC HORIZON2020 777119 InSilc, and MESTD, Republic of Serbia (projects III41007 and ON174028).

*Corresponding author, e-mail: markovac85@kg.ac.rs

P4

Modelling of Monocytes Behaviour inside the Bioreactor

Milica Nikolic[1,2,*], Tijana Sustersic[1,2,3], Igor Saveljic[1,3], Nihal Engin Vrana[4], and Nenad Filipovic[1,2,3]

[1] *Faculty of Engineering, University of Kragujevac, FINK, Sestre Janjic 6, Kragujevac, Serbia*
[2] *Steinbeis Advanced Risk Technologies Institute doo Kragujevac, SARTIK, Kneza Milosa 25, Kragujevac, Serbia*
[3] *Bioengineering Research and Development Center, BioIRC, Prvoslava Stojanovic 6, Kragujevac, Serbia*
[4] *PROTIP SAS, 8 Place de l hopital, Strasbourg, France*

Abstract:
Geometry of the bioreactor was created on the basis of the existing literature. Flow of the fluid, containing the monocyte cells, through the bioreactor was analysed. The model was created with open solver PAK and solved with the finite element method, as well as with commercial software ANSYS. The obtained results were compared between the solvers and compared to the experimental measurements. Good agreement of results was achieved.

Elements of the model were created as 3D 8-node elements. The fluid inside the bioreactor was modelled with the Navier-Stokes equation together with the continuity equation. Inside the fluid, a certain amount of the monocytes was taken into account. The number of monocyte cells was chosen from the literature. Fluid velocity at the inlet was prescribed, as well as open end, meaning zero pressure, at the outlet boundaries. Initially, monocytes concentration was assumed to be zero (empty bioreactor) and monocytes were introduced together with the fluid, so change in concentration was observed.

Movement of the monocyte cells was modelled as mass transport through the fluid. Monocytes are driven by drag force from fluid velocity, according to Stoke's law. Change in concentration of monocyte cells was monitored during the time for which the fluid flows through the bioreactor continuously, thanks to the peristaltic pump. The trajectories of the monocytes were obtained by using the calculated drag force from the fluid.

The developed model of bioreactor was used for simulation of the monocyte cells movement inside the bioreactor. Trajectories and concentration of the monocytes over time are in good agreement with the experimental measurements and the results obtained with two different finite element solvers, PAK and ANSYS, matched as well. The presented model of bioreactor and monocytes is a good starting point in further analysis of the monocytes. The next model will include modelling of a chemical reaction and binding of monocytes to the bottom of the bioreactor, in order to present the inflammatory process and reaction of the immune system.

Keywords:
bioreactor, fluid flow, monocyte cell, finite element modelling

This work has been funded by MESTD, Republic of Serbia (projects III41007 and 174028) and by the European Project H2020 PANBioRA 760921-2.

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

P10

PalFin: A Software Tool to Identify Specific Palindrome Motifs in mtDNA

Dinko Osmanković [1], Semir Dorić [2*], Lukić-Bilela Lada [3] and Naris Pojskić [2]

[1] *Faculty of Electrical Engineering, University of Sarajevo, Zmaja od Bosne bb. Kampus, 71 000 Sarajevo, Bosnia and Herzegovina*
[2] *Institute for Genetic Engineering and Biotechnology (INGEB), University of Sarajevo, Zmaja od Bosne 8 Kampus, 71 000 Sarajevo, Bosnia and Herzegovina*
[3] *Department of Biology, Faculty of Science, University of Sarajevo, Zmaja od Bosne 33-35, 71 000 Sarajevo, Bosnia and Herzegovina*

Abstract:
The occurrence of repetitive and/or palindromic elements in mitochondrial DNA (mtDNA) has been described across the Domain Eukarya. It has been reported that repetitive hairpin-forming elements could be involved in mtDNA rearrangements and replication of mtDNA. Also, they could play an important role in regulation of transcription as target sequences in interactions with DNA-binding proteins (transcription factors).

Due to the fact that poriferan mtDNA have retained some ancestral features, in particular multiple, relatively long non-coding regions with plenty of direct and inverted repeats (IR) and palindromes, a software tool for intergenic palindrome elements search and identification was created by using sequences from 62 sponge species as input data.

In this paper we present a novel algorithm (PalFin) for the extraction of such palindromic sequences. The algorithm is based on subsequence extraction and data analysis for which sequences from NCBI database were employed.

PalFin enables users an easier and more efficient search and identification of repetitive and palindromic elements aiming to elucidate the role of such elements in transcriptional regulation, especially in mitochondrial genomes of diploblastic animals.

Keywords:
PalFin, software, palindrome motif, Porifera, mtDNA, NCBI database

This work has been funded by University of Sarajevo, Institute for Genetic Engineering and Biotechnology.

*Corresponding author, e-mail: semir.doric@ingeb.unsa.ba

P1

Positional Biases of the Experimentally Characterized T-cell Epitopes

Mirjana D. Pavlović[1], Ana M. Jelović[2], Davorka R. Jandrlić[3], and Nenad S. Mitić[4*]

[1] *University of Belgrade, Institute of General and Physical Chemistry, Studentski Trg 12, Belgrade, Serbia*
[2] *Faculty of Transport and Traffic Engineering, University of Belgrade, 305 Vojvode Stepe, 11000 Belgrade, Serbia*
[3] *Faculty of Mechanical Engineering, University of Belgrade, Kraljice Marije 16, Belgrade, Serbia*
[4] *Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*

Abstract:
Large scale mapping of the experimentally validated T-cell epitopes on their antigens of origin had shown both similarity and differences in the positional biases and hydrophobicity of epitopes. T-cell epitopes, hosted in the Immune Epitope Database (IEDB), were captured and both positive and negative ones were analyzed. Epitopes were divided in 2 different datasets: a) epitopes tested for binding to HLA class-I or class II molecules and b) T-cell functional assay positive/negative (immunogenic/non-immunogenic) epitopes (with known HLA specificity). The percentage of epitopes, binding to either HLA-I or HLA-II molecules, and epitope frequency per region length, are much higher in the consensus of structurally stable (ordered) regions, defined by different disorder predictors (or majority of these predictors), than in the consensus of structurally unstable (disordered) regions. HLA-I binding epitopes are more frequent than HLA-II binders in ordered and mixed and/or boundary regions than in disordered regions. Epitopes, negative for binding to HLA-I molecules (HLA-I non-binders), were concentrated in ordered regions, being several times more frequent than HLA-I binders, while the frequency of HLA-II binders is significantly higher than non-binders in all types of predicted structural regions. Regarding T cell assay negative or positive epitopes (all HLA binding positive), HLA-I binders are preferentially concentrated in ordered regions and there is higher percentage of T cell assay negative than positive epitopes. However, among HLA-II binders, the relation between T cell assay negative and positive epitopes is almost equal, although the percentage of T cell assay positive epitopes is slightly higher than those of negative epitopes in all types of structural regions. These results could possibly be attributed to the overestimated prediction of HLA-I binding epitopes towards structurally ordered regions. Although T cell epitopes have positional bias to ordered regions, which are more hydrophobic than disordered regions, all classes of analyzed epitopes have higher number of epitopes with majority of hydrophilic amino acids. However, HLA-I binding epitopes (of various length) which are positive in T-cell functional assays are more hydrophobic than those that are negative in T-cell functional assays. The former result could be useful in better prediction of immunogenic epitopes.

Keywords:
T-cell epitopes, IEDB, disorder/order, HLA-binders

This work has been funded by MESTD, Republic of Serbia (projects 174021 and 174002).

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

P1

Video-based extraction of movement artifacts in electrogastrography signal

Nenad B. Popović[1*], Nadica Miljković[1], and Veljko Papić[1]

[1] *University of Belgrade – School of Electrical Engineering, Bulevar kralja Aleksandra 73, 11120 Belgrade, Serbia*

Abstract:
Electrogastrography (EGG) is non-invasive diagnostic technique for recording myoelectrical activity of stomach using cutaneous electrodes placed on the abdomen. Amplitude of EGG signal ranges from 100 μV to 500 μV with approximate frequency of 3 cycles per minute (cpm) which makes it highly vulnerable to artifacts originating from body movement. This paper presents a video-based analysis for movement artifacts detection during EGG recording and corresponding artifact rejection in EGG signal.

EGG was obtained simultaneously with video recording (90 minutes duration) in healthy volunteer (male, 25 years, 93 kg, 180 cm) in supine position after a test meal (commercially available oatmeal - 274 kcal).

In order to detect movements, new video sequence was generated by subtracting mean frame from original frames. Resulting frames had non-zero values only in pixels that changed during time caused by movements. Further, 1D signal was constructed by summation of absolute pixel values of each frame. Threshold (15% of maximum value) method was used to detect excessive changes in summed absolute pixels corresponding to movements. Sequences of EGG signals with detected movements were deleted from the final cleared sequence. In order to evaluate proposed method, percentages of spectral power share in normogastric range (2-4 cpm) were calculated for EGG signal with and without movement sequences. In addition, in order to suppress artifacts originating from electrocardiography (ECG) and breathing artifacts, EGG was filtered with Butterworth 3rd order band pass filter with cut off frequencies of 0.03 Hz and 0.25 Hz.

Spectral power shares in normogastric range for EGG signal were 33% and 40% before and after movement artifact removal, respectively. With this threshold only 247 samples (123.5 s) were deleted from the original signal.

Increase in spectral power share suggests that higher amount of valuable information is provided from cleared EGG signal. Future research should include optimization of proposed method and combination with other techniques for movement detection (e.g. accelerometers, electromyographic signals, and matched filter).

Keywords:
electrogastrography, video analysis, movement artifact

This work has been partially funded by MESTD, Republic of Serbia (project 175016).

*Corresponding author, e-mail: nenad.pop92@gmail.com

P1

3D Simulation of Inflammatory Process in Coronary Arteries

Igor Saveljic[1,2*], Themis Exarchos[3], Oberdan Parodi[4] and Nenad Filipovic[1,2]

[1] *Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia*
[2] *Bioengineering Research and Development Center, 34000 Kragujevac, Serbia*
[3] *Foundation of Research and Technology Hellas -Biomedical Research Institute, Ioannina, Greece, University of Ioannina, Ioannina, Greece*
[4] *Instituto di Fisiologia Clinica, Consiglio Nazionale delle Ricerche (IFC CNR), Pisa, Italy*

Abstract:
Atherosclerosis is a common dangerous disease in many developed countries. The inflammatory process of atherosclerosis leads to the formation of an atherosclerotic plaque in the intima layer of the blood vessel. This plaque changes geometry of the blood vessel's lumen by narrowing it and interacts with the blood flow. As the plaque continues to grow, the wall shear stress increases through the decreasing cross sections. The aim of this paper is to examine the influence of wall shear stress on the atherosclerosis development. The Navier-Stokes equations govern the blood motion in the lumen, the Darcy law is used for model blood filtration, Kedem-Katchalsky equations for the solute and flux exchanges between the lumen and the intima. The system of three additional reaction-diffusion equations that models the inflammatory process and lesion growth model in the intima was used. Determination of plaque location with computer simulation for a specific patient shows a potential benefit for prediction of disease progression.

Keywords:
atherosclerosis, computer simulation, plaque propagation, wall shear stress

This work has been funded by EC HORIZON2020 689068 SMARTool, and MESTD, Republic of Serbia (projects III41007, OI174028).

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

P10

Combined *in silico* and experimental approach to identify the peptide mimetic of the nanobody that stabilize functional conformational state of the beta2 adrenergic receptor (β2AR)

Milan Sencanski[1], Milka Vrecl[2*], Nevena Veljkovic[2], Sanja Glisic[2]

[1] *Center for Multidisciplinary Research, Institute of Nuclear Sciences VINCA, University of Belgrade, Mike Petrovića Alasa 14-16, 11001 Belgrade, Serbia*
[2] *Institute of Preclinical Sciences, Veterinary Faculty, University of Ljubljana, Slovenia*

Abstract:
Stabilization of specific G-protein coupled receptor (GPCR) conformation is achieved by ligand binding to orthosteric or allosteric sites on a GPCRs. A crucial unresolved issue in GPCRs activation/signaling is the role of receptor structural conformations in G protein/effector protein selection. One of the possible approaches to get comprehensive depiction of GPCRs activation dynamics are molecular simulations and recently described nanobody-derived intrabodies. Monomeric single-domain antibody (nanobody) from the Camelid family was found to allosterically bind to and stabilizes distinct conformational states of the β2AR. By applying informational spectrum method (ISM), a virtual spectroscopy method for investigation of the protein-protein interactions, we have designed peptide mimetic of the nanobody related to the β2AR (nanobody derived peptide, NDP). Further, interaction between NDP and the ligand-bound β2AR active conformation have been studied by protein-peptide docking, molecular dynamics simulations and metadynamics calculations of free energy binding. Finally, the affinity of selected NDPs towards agonist-activated β2AR was also studied by microscale thermophoresis (MST) and by bioluminescence resonance energy transfer (BRET) based β-arrestin 2 recruitment assay. MST data predicted micromolar range interaction of selected NDPs with the β2AR, while the preliminary β-arrestin 2 recruitment results suggest prospective further modification and optimization of NDPs toward effective modulators of the β2AR.

Keywords:
bioinformatics, informational spectrum method, molecular dynamics simulations, nanobody derived peptides, protein-protein interactions

This work has been funded by MESTD, Republic of Serbia (project 173001)

*Corresponding author, e-mail: milka.vrecl@vf.uni-lj.si

P4

Exploring the usefulness of graph properties in protein protein interaction predictions

Neven Šumonja, Nevena Veljković and Vladimir Perović*

Center for Multidisciplinary Research and Engineering, Vinca Institute of Nuclear Sciences, Mike Petrovića Alasa 12-14, Belgrade, Serbia

Abstract:
A proteome-scale map of the protein–protein interactions (PPI) enables a global view of cellular function and organization. However, experimentally validated PPI information is sparse and biased in favour of well-studied proteins and therefore, computational prediction methods add a value by prioritizing candidates and covering unexplored zones of the interactome. In this work we tested graph measures commonly used in network analysis as a new type of numerical features for encoding. In our model each protein is perceived as a node in PPI network and it is numerically encoded by selected graph properties. The encouraging evaluation results on human and yeast test sets suggest that this distinct class of attributes have considerable potential to improve methods that utilize machine learning algorithms and perform binary classification on interacting and non-interacting protein pairs

Keywords:
bioinformatics, graph features, protein-protein interactions

This work has been funded by MESTD, Republic of Serbia (project 173001).

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

P7

Numerical simulation of electrospinning using PAK and ANSYS software

Tijana Šušteršič[1,2,3*], Liliana Liverani[4], Aldo R. Boccaccini[4], Slobodan Savić[1] and Nenad Filipović[1,2,3]

[1] *Faculty of Engineering (FINK), University of Kragujevac, Sestre Janjic 6, Kragujevac, Serbia*
[2] *Bioengineering Research and Development Center (BioIRC), Prvoslava Stojanovica 6, Kragujevac, Serbia*
[3] *Steinbeis Advanced Risk Technologies Institute doo Kragujevac (SARTIK), Kneza Milosa 25, Kragujevac, Serbia*
[4] *Institute of Biomaterials, Department of Materials Science and Engineering, University of Erlangen-Nuremberg, Cauerstr. 6, Erlangen, Germany*

Abstract:
During the process of electrospinning, polymer solution is accelerated from a capillary orifice of the syringe under the influence of electric field, and the liquid jet is injected in the air that travels towards the grounded plate. Consequently, it is crucial to simulate the electrospinning process including the interaction of electric field and polymer. The motivation for this research was to analyze the effects of different solution, process and ambient parameters without the necessity to perform the experiments each time. Additionally, validated simulations in PAK software, developed at the University of Kragujevac, would create possibilities to include additional modifications and parameters that influence the process of electrospinning.

The simulations were performed using the Finite Element Method (FEM) based on geometry values obtained from commercially available EC-CLI device (IME Technologies, Geldrop, The Netherlands). Electrospinning parameters used in the experiment were adopted for simulation - 21G needle type; voltage pairs 15kV applied on the nozzle and 0kV on the collector; 0.8 ml/h flow rate. Material properties for 10wt% PVA solution were adopted from the literature values, as they could not be determined experimentally at this point. Laminar flow was assumed and simulation included flow field and electric field interaction.

The simulation results in PAK software show good agreement with commercially available software ANSYS and similar jet shapes were obtained during the simulations. Initial results also confirm the hypothesis that the jet shape during electrospinning can indicate whether setting of the chosen electrospinning parameters would result in good quality fibers. Future studies would include parameter modifications in order to examine jet shapes when optimal and non-optimal parameters are used. The ultimate goal is to provide users with stand-alone application for finding optimal parameters in electrospinning, which would reduce time-consuming process of performing experiments until relevant parameters are determined. The beneficial effect of such simulations could also be seen in reduced solution material consumption, maintenance costs etc.

Keywords:
electrospinning, PAK software, computational simulation, finite element method (FEM)

This work has been funded by MESTD, Republic of Serbia (project III41007 and 174028), European Project H2020 PANBioRA [grant number 760921-2] and KMM-VIN Research Fellowship 2017 (Call 9).

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

P2

Application of bioinformatics in the identification of autohtonous bacterial strains of Vranjska Banja thermal springs based on different methods

Nemanja Tasić[1], Srđan Tasić[2] and Irena Tasić[3]

[1] *Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia*
[2] *Higher School of Applied Professional Studies, Filipa Filipovića 20, 17500 Vranje, Serbia*
[3] *The College of academic studies "Dositej", Nušićeva 12a, 11000 Belgrade, Serbia*

Abstract:
Three strains of thermophilic bacteria (A70, C70 and ST2) isolated from the thermal springs in Vranjska Banja were the subject of the molecular and biochemical characterisation. With a water temperature ranging from 80 – 96 0C, these springs, along with Iceland geysers, are the hottest in Europe. In these sodium-hydrocarbonate-sulphate and sulphide hyperthermal waters the total β-activity is 475.8 mBq/L and α-activity is •37.0 Bq/L. Standard microbiological methods were used for the bacteria isolation. The biochemical characterisation was performed using the API 50 CHB system (bioMérieux) and APIWEBTM software Ver. 4.1. A70 and C70 strains were characterised as Geobacillus stearothermophilus (%ID = 99.9 T=0.36 and %ID = 99.7 T=0.54, respectively), while the ST2 strain was identified as Bacillus pumilus (%ID = 99.9 T=0.45).

Molecular characterisation of these three strains was performed by analysis of the tuf gene, which encodes the elongation factor Tu. Both strands of the PCR fragments were sequenced with the same primers used for amplification. The DNA sequences were compared to those deposited in data bases using the BLAST program (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn). The highest level of similarity was determined between the A70 and the KCTC 3570 Geobacillus thermoleovorans (98% identity), between the C70 strain and the PS11 Geobacillus stearothermophilus (98% identity), as well as between the ST2 strain and the ATCC 9789 Bacillus licheniformis (98% identity). Only in case of the C70 strain, both approaches (biochemical and molecular characterisation) identified the strain as Geobacillus stearothermophilus. Other two strains were classified in the same genus by biochemical characterization, while at molecular level they were identify as different species.

Given that all the conducted analyses yielded a substantial number of data, they were processed and compared using biostatistics methods and tools in order to achieve the highest probability of resulted taxonomic classification. Modern research contributes to the analysis of a significant number of variables which is why considerably more statistical analyses are involved in their interpretation and presentation.

Our results indicate that different methods are needed for proper determination and characterisation of isolates/strains. Regarding taxonomy, molecular methods are the most precise, while for physiological specificity biochemical methods are more reliable.

Keywords:
thermal spring, native bacteria, bioinformatics

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

P2

Determination of hTERT promoter methylation status using methylation specific PCR

Jovanka Trifunovic

Faculty of Pharmacy and Health, University of Travnik, Travnik, Bosnia and Herzegovina

Abstract:
DNK methylation is significant for process that participate in embryonic development,cell differentiation,maintenance of homeostasis in the cells. The regulation of gene expression and maintenance of genome stability are mechanisms that ensure that these processes are carried out.

Renal Cell Carcinoma (RCC) is the most common renal cancer in adults and includes several subtypes that may be distinguished by their histology, genetic background, clinical course and treatment. Human telomerase reverse transcriptase (hTERT) is a crucial enzyme for telomere maintenance, has been linked to RCC development. The aims of this study were to search for genetic and epigenetic alterations in hTERT methylation and to establish a possible association between molecular and clinico-pathological characteristics of RCC.

DNA was extracted from 31 formalin-fixed, paraffin-embedded tumor samples and 23 blood samples from 54 patients with RCC. Polymerase chain reaction (PCR) products were sequenced and analyzed using the Sequencher software. The hTERT promoter methylation status was determined by methylation specific PCR (MSP). In order to detect DNA sequence alterations after bisulphite treatment, MSP was used to distinguish methylated from unmethylated alleles. Genomic DNA from lymphocytes of healthy donors was used as control for unmethylated genes and the same DNA, treated in vitro with SssI methyltransferase, was used as control for methylated genes.

HTERT promoter was methylated in 17 of the 31 tumor samples (54.8%). Interestingly, in 71% of papillary carcinomas the promoter was unmethylated, wereas in 100% of chromophobe carcinomas it was methylated. An association was established between methylation and histological type of RCC (p=0.047).

In conclusion, hTERT, via its promoter methylation seems to play a role in renal cell carcinoma biology. hTERT promoter methylation status is related to RCC histology.

Keywords:
hTERT methylation, methylation specific PCR, RCC

This work has been funded by MESTD, Republic of Serbia (project 175059).

*Corresponding author, e-mail: drjotrifunovic@aol.com

P3

Numerical Simulation of Blood Flow and Plaque Progression in Right Femoral Artery Bypass Patient-Specific Case

Ana Vulovic[1,2*], Danko Milasinovic[3], Dragan Sekulic[4] and Nenad Filipovic[1,2]

[1*] *Faculty of Engineering, University of Kragujevac, Sestre Janjic 6, 34000 Kragujevac, Serbia*
[2] *Bioengineering Research and Development Center (BioIRC), Prvoslava Stojanovica 6, Kragujevac, Serbia*
[3] *Faculty of Hotel Management and Tourism, University of Kragujevac, Vojvodjanska bb, 36210 Vrnjacka Banja, Serbia*
[4] *Military Medical Academy, Clinic for Vascular and Endovascular Surgery, Crnotravska 17, 11000 Belgrade, Serbia*

Abstract:
The aim of this study was to compare blood flow (pressures, shear stress, velocities) and plaque progression patterns in a diseased and by-passed right femoral artery (FA) using the finite element method. The patient's geometry of the diseased FA was obtained from the corresponding computer tomography (CT) medical scans, which showed serious diameter narrowing. The by-passed FA was obtained virtually, using software for 3D visualization from images, and software for 3D modeling to create the by-pass. The mass transfer within the blood lumen and through the arterial wall was coupled with the blood flow and modeled by convection-diffusion equation. Blood flow through the obtained 3D model was governed by Navier-Stokes equations including the continuity equation. The transport of low-density lipoprotein (LDL) in the vessel lumen was described by Kedem-Katchalsky equations. The main advantage of the proposed procedure is reduced invasiveness and time efficiency, while enabling the user-friendly three-dimensional visualization of objects obtained from the patient's CT scans. Therefore, it is suitable for monitoring the patient's condition, potential medical procedures risk assessment and decision on medical therapy.

Keywords:
femoral arteries, finite element method, femoral bypass, plaque growth, shear stress

This work has been funded by the grants: EC HORIZON2020 689068 and SMARToolMESTD, and MESTD, Republic of Serbia (project III41007 and 174028)

*Corresponding author, e-mail: ana.vulovic995@gmail.com

P5

Development of anatomically correct human mandible finite element model from CT-scans

Ksenija Zelic Mihajlovic[1], Arso Vukicevic[2], Petar Milovanovic[1], Marija Đurić[1], Nenad Filipovic[2,3]

[1] *Laboratory for Anthropology, Institute of Anatomy, School of Medicine, University of Belgrade, Dr Subotića 4/2, 11000 Belgrade, Serbia*
[2] *Faculty of Engineering, University of Kragujevac, 6 Sestre Janjic Str., 34000 Kragujevac, Serbia*
[3] *Bioengineering Research and Development Center Kragujevac, Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia*

Abstract:
Background and aim For any computational analysis to be carried out on human jaws there is a need for detailed model with all anatomical and morphological characteristics of the real jaws. The aim of this article was to propose the procedure for development of finite element model of human mandible with present teeth (with exception of third molars).

Methods Dry mandible (from archeological sample) was used. No visible bone loss was present and all teeth except third molars were present. The imaging was obtained by Cone Beam Computed Tomography-CBCT (the jaw), and Micro Computed Tomography-Micro CT (each tooth separately). All the scans were stored in the standard DICOM format for the further analysis. For each tooth, we considered its enamel, dentin, pulp chamber, and periodontal ligament (PDL), while the root cementum was neglected. For the bone we considered trabecular bone and cortical bone. Mimics software version 10 (Materialise, Leuven, Belgium) was used for the reconstruction of the FE models from the CBCT scans. In order to optimize the quality of the triangle meshes for the further FEA, we used the REMESH module attached to Mimics software. At last, by using Geomagic Studio 10 software (Geomagic GmbH, Stuttgart, Germany) we assembled the extracted parts into the models. A PDL as the 200 μm-thick shell was additionally generated. For every part of the model, defined with a single STL file, a 3D volume discretization was carried out using TetGen meshing software (Hang Si, WIAS, Berlin, Germany). At this stage, a high-quality four-nodal tetrahedral elements (TET4) mesh was. After splitting the tetrahedral elements into eight-nodal hexahedral the values of the Young's moduli and the Poisson's ratios for dental tissues, PDL, cortical and trabecular bone were taken from the literature.

We gained very detailed mandible model containing teeth with all parts present. The stress distribution (Von Mises, compressive, tensile) and effective displacements during occlusal forces were analyzed in order to verify the quality of the model.

Keywords:
bioengineering, human mandible model, CT scan

This work has been funded by MESTD, Republic of Serbia (project 45005).

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

P6

Optimization of Electrochemical Parameters for Detection of microRNA: Computer Simulation and Experimental Study

Marko Živanović[1,2*] and Nenad Filipović[2,3]

[1] *University of Kragujevac, Faculty of Science, Radoja Domanovića 12, 34000 Kragujevac, Serbia*
[2] *BioIRC, Bioengineering R&D Center, Prvoslava Stojanovića 6, 34000 Kragujevac, Serbia*
[3] *University of Kragujevac, Faculty of Engineering, Sestre Janjić 6, 34000 Kragujevac, Serbia*

Abstract:
The use of electrochemical methods in the examination of the structure, function and detection of biomacromolecules is widespread. For precise electrochemical determination of biomacromolecules, it is necessary to optimize the experimental conditions. Optimization is a timely and experimentally required process because it is necessary to match many parameters to get precise, accurate and reproducible results. microRNAs could be considered as markers of any healthy or unhealthy cell state. Thus, microRNAs could be considered as biomarkers for cancer detection. Namely, in addition to many other microRNAs that are constantly expressed in both healthy and tumor cells, we have chosen one of the most characteristic – microRNA21, whose concentration was significantly increased in patients suffering from breast cancer.

In our experiment, we used phosphate buffer of various pHs, ionic strengths, with or without salt addition, in temperature-controlled conditions with stirring in different speeds. Each of these parameters must be individually examined and aligned with the rest. Often the combination of several parameters of determination results in a special synergy, which is difficult to predict at the beginning of experimental process. Therefore, in addition to experimental electrochemical examination, we also used advanced computational methods to predict the numerical model for a given parameter or a combination of parameters that would affect the expression of the expected electrochemical signal.

In the modeling of electrochemical deposition on the electrode we used Nernst-Planck definition of species flux through homogeneous media. The flux resulting from the electrochemical potential gradient is typically decoupled into a diffusion term corresponding to activity or concentration gradient driven flow and an electromigration term that accounts for the force of the electric field on charged molecules.

The use of computational methods in electrochemistry enabled us the save of time and resources in everyday work. In this sense, *in-silico* optimization is the first step in our work. In the future, we will try to develop advanced mathematical models that accurately describe the effects of electrochemical deposition to the electrode.

Keywords:
Computer Simulation, Electrochemistry, Optimization.

This work has been funded by MESTD, Republic of Serbia (project III41007) and PanBioRA- Horizon 2020 research and innovation programme, No 760921.

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

P6

FINDbase: worldwide database for clinically relevant genomic variation allele frequencies

Branka Zukic[1], Biljana Stankovic[1], Nikola Kotur[1], Vladimir Gasic[1], Kristel Klaassen[1], Ksenija Vukicevic[2], George Patrinos[3], and Sonja Pavlovic[1]

[1] *Laboratory for Molecular Biomedicine, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11010 Belgrade, Serbia*
[2] *Department of Physiology, Faculty of Medical Sciences, University of Kragujevac, Svetozara Markovica 69, 34000 Kragujevac, Serbia*
[3] *Department of Pharmacy, School of Health Sciences, University of Patras, University Campus Rion, GR-26504, Patras, Greece*

Abstract:
FINDbase (Frequency of INherited Disorders database, http://www.findbase.org) is a comprehensive online data repository that in a structured manner records the prevalence of clinically relevant genomic variants in various populations worldwide, such as pathogenic variants leading mostly to monogenic disorders and pharmacogenomics biomarkers, all in well-distinct data modules. The incidence of rare genetic diseases in various populations is recorded also. FINDbase is freely available and no registration is needed for data querying.

The data came from previously published reports as well as from unpublished information contributed from individual researchers prior of publication. The contributor's unique ResearcherID follows the microattribution approach, allowing unambiguous identification of curated data.

Serbian research team has actively participated in all phases of design and development of the FINDbase.

Causative Mutations module of FINDbase documents the frequencies of causative mutations leading to inherited disorders in various populations worldwide. Database records include the population, the ethnic group and/or the geographic region, the gene name and its variation parameters, the rare allele frequencies, linked to the respective Online Mendelian Inheritance in Man (OMIM) and the Human Gene Mutation database (HGMD) entries. This module includes data for more than 3,800 disease-causing mutations across 26 genes, representing over 100,000 individuals from 92 populations.

Pharmacogenomic Markers module of FINDbase documents the frequencies of pharmacogenetically relevant SNVs in various populations worldwide. Database records include the population, the ethnic group and/or the geographic region, the gene name and its variation parameters, the rare allele frequencies, linked to the respective OMIM and the PharmGKB entries. This module includes data for 144 pharmacogenomic markers across 14 genes, representing 87,000 individuals from 150 populations. Interrelating the PGx data module with DruGeVar database (http://drugevar.genomicmedicinealliance.org) provide users with the best of both resources.

The National/Ethnic Genetic Databases (NEGDB) modules records information over the described genetic heterogeneity of an ethnic groups/populations. FINDbase comprises 90 individual ETHNOS-based NEGDBs in all 5 continents.

In the era of big data, FINDbase is introducing a system for microarray or next-generation sequencing records uploading and creating an algorithm to automatically calculate clinically relevant genomic variants in an aggregated manner to ensure data anonymity.

Keywords:
FINDbase, microattribution approach, Causative Mutations module, Pharmacogenomic Markers module, National/Ethnic Genetic Databases

This work has been funded by MESTD, Republic of Serbia (project III41004).

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

MAIN ORGANIZER

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade

MAIN CO-ORGANIZERS

Vinča Institute of Nuclear Sciences,
University of Belgrade

Faculty of Mathematics,
University of Belgrade

Mathematical Institute
of SASA, Belgrade

Faculty of Biology, University of
Belgrade

CO-ORGANIZERS

Serbian Genetic Society

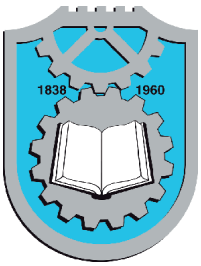
Faculty of Agriculture,
University of Belgrade

Institute for Medical Research,
University of Belgrade

Faculty of Sciences,
University of Novi Sad

Faculty of Chemistry,
University of Belgrade

Institute for Biological Research
"Siniša Stanković", Belgrade



Faculty of Engineering,
University of Kragujevac

Institute for General and Physical
Chemistry, Belgrade

Faculty of Sciences and Mathematics,
University of Niš

Maize Research Institute Zemun Polje

BioIRC – Bioengineering Research and
Development Center

COST European Cooperation in Science
and Technology

SERBIAN SOCIETY FOR BIOINFORMATICS AND
COMPUTATIONAL BIOLOGY

BELGRADE BIOINFORMATICS CONFERENCE 2018 HAS BEEN SUPPORTED BY

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

The Abdus Salam International Centre for Theoretical Physics

BELGRADE BIOINFORMATICS CONFERENCE 2018 HAS BEEN SPONSORED BY

PLATINUM SPONSOR (TB)

Telekom Srbija

vivogen® BIOSISTEMI GRUPA

GOLD SPONSOR (GB)

QIAGEN

SILVER SPONSOR (MG)

ALFA GENETICS

KEFO® SINCE 1949

Labena

BRONZE SPONSOR (KB)

DSP Chromatography

Agilent Technologies Authorized Distributor

ProMedia Specialists In Laboratory Supplying

RTC

INDEX