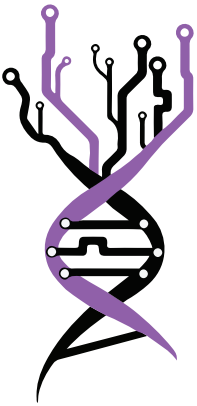
#BelBi2023 • Belgrade, Serbia

BOOK OF ABSTRACTS



4th Belgrade Bioinformatics Conference

HYBRID • 19 - 23 JUNE **2023**

EDITORS
Dr. Ivana Morić
Dr. Valentina Đorđević

ISBN: 978-86-82679-14-1

belbi.bg.ac.rs

Title 4th Belgrade Bioinformatics Conference
BOOK OF ABSTRACTS

Publisher Institute of Molecular Genetics and Genetic Engineering, University of Belgrade
Vojvode Stepe 444a, Belgrade, Serbia
https://www.imgge.bg.ac.rs/

Editors dr. Ivana Morić
dr. Valentina Đorđevic

Technical editor Dušan Radojević

ISBN **978-86-82679-14-1**

BelBi2023 Committees

International Advisory Board

Alessandro Treves,
International School for Advanced Studies, Trieste, Italy

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

Konstantin Severinov,
Waksman Institute for Microbiology, Rutgers University, Piscataway, USA

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

Milena Stevanović,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia

Oxana Galzitskaya,
Institute of Protein Research, Russian Academy of Sciences, Moscow, Russia

Paul Sorba,
Laboratory of Theoretical Physics and CNRS, Annecy, France

Peter Tompa,
VIB Structural Biology Research Center, Flanders Institute for Biotechnology, Brussels, Belgium

Predrag Radivojac,
Khoury College of Computer Sciences, Northeastern University, Boston, SA

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

Vladimir Brusić,
Li Dak Sum Chair Professor of Computer Science, University of Nottingham, Ningbo, China

Vladimir Uversky,
Department of Molecular Medicine, University of South Florida, Tampa, USA

Yuriy Orlov,
I.M. Sechenov First Moscow State Medical University, Moscow, Russia

Zoran Ognjanović,
Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia

International Program Committee

Alexandre de Brevern,
INSERM, Université Paris Cité, Université de la Réunion, Paris, France

Branislava Gemović,
VINCA Institute of Nuclear Sciences, University of Belgrade, Serbia

Branka Zukić,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia

Branko Dragovich,
Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade, Serbia

Dragan Matić,
Faculty of Sciences, Department of Mathematics and Informatics, University of Banja Luka, Bosnia and Herzegovina

George Patrinos,
Department of Pharmacy, University of Patras, Greece

Gordana Pavlović-Lazetić,
Faculty of Mathematics, University of Belgrade, Serbia

Hong-Yu OU,
Shanghai Jiao Tong University, The Microbial Bioinformatics Group, State Key Laboratory of Microbial Metabolism, Shanghai, China

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

Jelena Bojović,
Center for the Fourth Industrial Revolution, Belgrade, Serbia

Marko Đordević,
Faculty of Biology, University of Belgrade, Serbia

Mirjana Maljković,
Faculty of Mathematics,
University of Belgrade, Serbia

Nataša Kovacevic-Grujičić,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia

Nataša Pržulj,
Catalan Institution for Research and Advanced Studies (ICREA), Spain;
Barcelona Supercomputing Center, Spain;
University College London, UK

Nenad Mitić,
Faculty of Mathematics,
University of Belgrade, Serbia

Nevena Veljković,
VINCA Institute of Nuclear Sciences,
University of Belgrade, Serbia

Predrag Radivojac,
Khoury College of Computer Sciences,
Northeastern University, Boston, USA

Saša Malkov,
Faculty of Mathematics,
University of Belgrade, Serbia

Sergei Kozyrev,
Department of Mathematical Physics,
Steklov Mathematical Institute RAS,
Moscow, Russia

Valentina Đorđević,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia

Vladimir Babenko,
Institute of Cytology and Genetics,
Novosibirsk, Russia

Local Organizing Committee

Anđela Rodić,
Faculty of Biology, University of Belgrade

Branislava Gemović,
VINCA Institute of Nuclear Sciences,
University of Belgrade

Branko Dragovich,
Mathematical Institute, Serbian Academy of Sciences and Arts, Belgrade

Gordana Pavlović-Lazetić,
Faculty of Mathematics,
University of Belgrade

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

Jovana Kovačević,
Faculty of Mathematics,
University of Belgrade

Marko Đorđević,
Faculty of Biology, University of Belgrade

Marko Živanović,
Bioengineering Research and Development Center

Nenad Filipović,
Faculty of Engineering, University of Kragujevac

Nenad Mitić,
Faculty of Mathematics,
University of Belgrade

Radoslav Davidović,
VINCA Institute of Nuclear Sciences,
University of Belgrade

Saša Malkov,
Faculty of Mathematics,
University of Belgrade

Valentina Đorđević,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

Željko D. Popović,
Faculty of Sciences, University of Novi Sad

FOREWORD

Dear colleagues and friends,

The 4th Belgrade Bioinformatics Conference - BelBi2023, where many high-quality scientific contributions were presented, has just ended. With great thanks to all participants, we now proudly present a book of abstracts that both reflects the scientific abundance and diversity of the conference and serves as a reminder of a memorable event.

Several research institutions, faculties, and scientific societies from Serbia joined forces in organizing this international conference, which covered numerous topics in computational biology, bioinformatics, and biomedical and health informatics. The main goal of BelBi2023 was to foster contact between scientists, both early stage career and senior researchers, allowing them to share experiences and latest advances in their fields. We sincerely hope that BelBi2023 has served as a platform for researchers from around the world to meet, initiate new collaborations, and expand professional contacts, and that all of you would become a part of the growing BelBi community.

We are grateful and proud to have welcomed more than 250 researchers from 21 countries. We have had 28 scientific sessions, consisting of more than 60 lectures (including eight Keynote talks), 47 presented posters, as well as three workshops and one satellite event – COST action. We have also organized seven industry lectures, including the NGS Challenge, two Meet the Expert Sessions, and one Business Coffee Break where ten start-up companies took part. And finally, the future BIO4 campus was presented and first panel on Serbia's resources for storage and analyses of genetic data was organized.

We would like to thank all the members of the International Advisory Board and the International Program Committee for their efforts and help in making this event a success. We are very grateful to the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, SAIGE project, and UNDP-Serbia for their support. Finally, the Local Organizing Committee is very grateful to all the sponsors of the conference - BGI, Illumina & Elta'90MS, PacBio & East Diagnostics, ThermoFisher Scientific & Vivogen, Huawei, Labena, DSP Chromatography, RNIDS, Telekom Srbija, Alfa Genetics, Kefo and Superlab, hoping that they will stay with us for many years to come.

Looking forward to seeing you again at the 5th Belgrade Bioinformatics Conference.

Belgrade, July 2023

Dr. Valentina Đorđević
& Dr. Ivana Morić,
On behalf of BelBi2023
Organizing Committee

ORGANIZER

Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

MAIN CO-ORGANIZERS

Faculty of Biology, University of Belgrade

Faculty of Mathematics, University of Belgrade

Mathematical Institute of SASA, Belgrade

Vinča Institute of Nuclear Sciences, University of Belgrade

Serbian Society for Bioinformatics and Computational Biology

CO-ORGANIZERS

Faculty of Sciences, University of Novi Sad

BioIRC - Bioengineering Research and Development Center

Faculty of Engineering, University of Kragujevac

Centre for the Fourth Industrial Revolution in Serbia

Serbian Society for Molecular Biology

Biophysical Society of Serbia

TABLE OF CONTENT

KEYNOTE LECTURES

INVITED LECTURES

ORAL PRESENTATION

POSTER PRESENTATION

Big Data in Biology: How EMBL delivers big data for biology, and some highlights of its application to human disease biology

Ewan Birney[1]

[1]European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom
birney@ebi.ac.uk

Molecular biology is now a leading example of a data intensive science, with both pragmatic and theoretical challenges being raised by data volumes and dimensionality of the data. These changes are present in both "large scale" consortia science and small scale science, and across now a broad range of applications – from human health, through to agriculture and ecosystems. All of molecular life science is feeling this effect. The European Molecular Biology Laboratory (EMBL) – Europe's only intergovernmental research organisation in the life sciences is at the forefront of these developments performing both excellent research and providing world leading services to enable science across Europe.
This shift in modality is creating a wealth of new opportunities and has some accompanying challenges. In particular there is a continued need for a robust information infrastructure for molecular biology. This ranges from the physical aspects of dealing with data volume through to the more statistically challenging aspects of interpreting it. A particular problem is finding causal relationships in the high level of correlative data. Genetic data are particular useful in resolving these issues. I will present how EMBL pursues this science and give examples from my own research that spans human genetics research through to partnering for clinical application.

A tale of two stories: data-driven precision medicine and precision public health

Kristel Van Steen[1,2]

[1] BIO3 Systems Genetics, GIGA-R, Université de Liège, 4000 Liège, Belgium
[2] BIO3 Systems Medicine, Department of Human Genetics, KU Leuven, 3000 Leuven, Belgium
kristel.vansteen@uliege.be

Big Data offers opportunities in health care to refine individuals' characterization and thus complement traditional precision medicine approaches toward individual-targeted prevention, diagnosis and treatment management. Not surprisingly, network theory plays a vital role in modelling Big Data: the higher the number of measurements, the higher the number of potential relationships or dependencies among them. Recent developments have shown the complementary value of personalizing population-based networks for individuals (Menche et al. 2017, Dimitrakopoulos et al. 2018) or deriving individual-specific networks via populations of cells (Gosak et al. 2018, Li et al. 2023).
Individual-specific networks do not necessarily require repeated measurements over time or in space. Reverse-engineered individual-specific networks (Kuijjer et al. 2019) from an aggregate network (hereafter referred to as ISNs) allow for investigating the impact of individual-level network wirings, paths or connectivity on medical decision-making in the individual's interest. Wondering about the utility of these ISNs, we illustrate by example from microbiome and gene co-expression experiments how ISNs give complementary insights in dynamic network biomarker identification and can reveal (genetic modifiers of) co-eQTLs as direct or indirect regulators of gene co-expression.

Keywords: individual-specific networks, precision medicine, precision public health, Big Data science

Acknowledgement: This work received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreements N° 813533 (mlfpm.eu) and N° 860895 (h2020transys.eu). We are grateful to all former and current BIO3 members for inspiring discussions.

Machine intelligence and network science for complex systems big data analysis

Carlo Vittorio Cannistraci[1]

[1] Center for Complex Network Intelligence, Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, 160 Chengfu Rd., Beijing, China
kalokagathos.agon@gmail.com

I will present our research at the Center for Complex Network Intelligence (CCNI) that I recently established in the Tsinghua Laboratory of Brain and Intelligence at the Tsinghua University in Beijing. We adopt a transdisciplinary approach integrating information theory, machine learning and network science to investigate the physics of adaptive complex networked systems at different scales, from molecules to ecological and social systems, with a particular attention to biology and medicine, and a new emerging interest for the analysis of complex big data in social and economic science.
Our theoretical effort is to translate advanced mathematical paradigms typically adopted in theoretical physics (such as topology, network and manifold theory) to characterize many-body interactions in complex systems. We apply the theoretical frameworks we invent in the mission to develop computational tools for machine intelligent systems and network analysis. We deal with: prediction of wiring in networks, sparse deep learning, network geometry and multiscale-combinatorial marker design for quantification of topological modifications in complex networks. This talk will focus on two main theoretical innovation. Firstly, the development of machine learning and computational solutions for network geometry, topological estimation of nonlinear relations in high-dimensional data (or in complex networks) and its relevance for applications in big data, with a emphasis on brain connectome analysis. Secondly, we will discuss the Local Community Paradigm (LCP) and its recent extension to the Cannistraci-Hebb network automata, which are brain-inspired theories proposed to model local-topology-dependent link-growth in complex networks and therefore are useful to devise topological methods for link prediction in sparse deep learning, or monopartite and bipartite networks, such as molecular drug-target interactions and product-consumer networks.

Keywords: Network topology and geometry, network automata, network biology, network neuroscience, artificial intelligence.

Acknowledgement: The author acknowledge all the collaborators and institutions that in years of research contributed to the research presented in the talk.

Targeting LLPS in disease: a new modality in drug development

Peter Tompa[1,2]
[1] VIB-VUB Center for Structural Biology, Brussels, Belgium
[2] Vrije Universiteit Brussel (VUB), Dept. DBIT, Brussels, Belgium
peter.tompa@vub.be

Biomolecular condensation is a process whereby many macromolecules (proteins and RNAs) form non-stoichiometric, functional assemblies. The dominant mechanism of such biomolecular condensation is liquid-liquid phase separation (LLPS), which leads to the formation of membraneless organelles (MLOs), such as the nucleolus and stress granules, in the cell. The proteins involved often have a high proportion of intrinsic structural disorder, which drive LLPS by transient, multivalent interactions. As MLOs play key roles in cell signaling, the misregulation of their formation and dissolution often leads to diseases termed "condensatopathies". In my presentation, I will outline the basic mechanisms leading to such disease states, focusing on cancer, viral infections and neurodegeneration. I will also discuss the different potential strategies for correcting these errors in cell signaling, and show through specific examples how drug candidates, "c-mods" capable of correcting MLO misregulation, can be developed.

Keywords: LLPS databases, LLPS mechanism, LLPS targeting, condensatopathy, ALS/FTD

Translating Bioinformatics Back To Healthcare: Facilitating the use of Artificial Intelligence at UW Medicine

Sean D. Mooney[1,2,3]

[1] Department of Biomedical Informatics and Medical Education,
University of Washington, 850 Republican St, Seattle, WA, 98109
[2] Institute for Medical Data Science,
University of Washington, 850 Republican St, Seattle, WA, 98109
[3] UW Medicine, 1001 4th Ave, Seattle, WA, 98154
sdmooney@uw.edu

It is an opportune time to be engaged in the research and application of informatics in biomedicine. The increased use of electronic and personal health records and personal mobile devices is creating many opportunities at research academic medical centers. At the University of Washington, I believe we are laying the groundwork to build the informatics and information technology infrastructure to support research on personalized approaches and the use of data science to enable them. We are beginning to see the early successes of these efforts and I will describe some of them. But there are many challenges, for example, we continue to generate massive amounts of data that is largely uncurated. This includes images, genomes and other -omics datasets, personal monitors, electronic health records, etc. In this presentation, I will discuss our support of data for research use within UW Medicine, our efforts to build new machine learning and data science approaches using clinical datasets, and our efforts to develop new machine learning methods and to implement them so that we can study the impacts of their use.

Keywords: bioinformatics, precision medicine, research computing, data mining, healthcare, genetics

Bioinformatics education course on gene networks reconstruction using online tools

Yuriy L. Orlov[1, 2, 3], Anastasia A. Anashkina[1]

[1] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia
[2] Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow 117198, Russia
[3] The Digital Health Institute, I.M.Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow 119991, Russia
y.orlov@sechenov.ru

Bioinformatics education requires the use of online computer tools for modeling protein-protein interactions, visual presentation of the networks, access to open databases. The usage of online bioinformatics tools makes it possible to reconstruct both protein and genes networks, and develop modeling skills for students. We consider the issues of computer reconstruction of gene networks - complexes of interacting macromolecules - using a list of genes associated with a particular disease, or a complex disorder based on public online bioinformatics tools - STRING-DB, GeneMANIA, Metascape, Cytoscape applications. Examples of computer reconstruction and visualization of gene networks of oncological diseases including glioma, breast cancer, as well as complex mental disorders such as Parkinson's disease, schizophrenia, were recently published in co-authorship with the students.

The use of only online bioinformatics tools is educational in nature, focused on students, both in mathematics and in natural sciences and medical disciplines, who do not have enough skills in computer science, programming, and writing their own code. Automatic construction of lists of genes associated with a disease using open databases (OMIM, GeneCards, MalaCards), computer reconstruction of gene networks, calculations of enrichment statistics for gene ontology categories have been successfully mastered by students. The educational bioinformatics materials designed for the students and with the students were tested at several universities in Russia, including courses in English for foreign students studied in Russia.

The tasks of digitalization of medicine, the development of IT technologies are in the priority. The epidemic situation that has existed in recent years and the forced transition to distance learning had accelerated the adoption of measures to change the formats of education, the emergence of new learning platforms. Note a number of qualitatively new tasks of education in the field of digital healthcare, such as the use of blockchain technologies, the use of Artificial Intelligence (AI) methods in support of medical decision-making. Overall, the educational course developed includes a theoretical part (video lessons) and a practical part - performing tasks on the use of computer programs and databases that have found a number of applications for medical problems in the reconstruction and analysis of networks of interactions of macromolecules.

Keywords: bioinformatics, education, data mining, computer science, gene networks

A New Framework for the Use of Variant Interpretation Tools in Clinical Practice

Predrag Radivojac[1]

[1] Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA
predrag@northeastern.edu

Current ACMG/AMP guidelines for the use of sequence variants for genetic diagnosis and treatment permit the use of in silico predictors as Supporting evidence (PP3 and BP4 criteria). These criteria, however, lack quantitative support and leave clinicians and scientists without standards for applying these criteria, leading to large interpretation variability. To address this challenge, our team built upon previous work and introduced a novel criterion that can be used to calibrate any computational model or any other continuous-scale evidence on any variant type. We used it to estimate score intervals corresponding to the four strengths of evidence for pathogenicity and benignity for fourteen missense variant interpretation tools on a carefully assembled data sets of known pathogenic and benign variants. We found that most tools achieved the Supporting evidence level for both pathogenic and benign classification using newly established data-driven thresholds. Importantly, at appropriate score thresholds, several in silico methods can also provide Moderate and Strong evidence levels for a limited number of variants. Based on these findings, we provided recommendations for quantitative revisions of the PP3 and BP4 criteria within ACMG/AMP guidelines and the future assessment of *in silico* methods for clinical interpretation.

Privacy-preserving Systems Medicine

Jan Baumbach[1,2]

[1] Institute for Computational Systems Biology, University of Hamburg, Notkestrasse 9, 22607 Hamburg, Germany
[2] Computational BioMedicine lab, Institute of Mathematics and Computer Science, University of Southern Denmark, Campusvej 1, 5000 Odense M, Denmark
Jan.baumbach@uni-hamburg.de

Artificial intelligence (AI) offers game-changing opportunities to healthcare. However, it also harbors risks to patient privacy in particular when dealing with sensitive clinical data stored in critical healthcare IT infrastructure. Specifically, data exchange over the internet is perceived insurmountable, posing a roadblock hampering big-data-based medical innovations.
We created a novel AI platform, the FeatureCloud AI app store that is based on the idea of federated learning where only model parameters are communicated. To maximize privacy, sensitive datasets remain stored locally and are analysed behind safe firewalls to assure the high standards in data privacy in order to (by design) comply with the strict GDPR.
We will exemplarly investigate the power of FeatureCloud apps for decentralized (1) genome-wide association studies (GWAS), (2) gene expression data mining, and (3) time-to-event data analytics to demonstrate how FeatureCloud may enhance worldwide collaboration, accelerate innovation, and democratize scientific data usage. We show that apps developed in FeatureCloud can produce highly similar results compared to centralized approaches and scale well for an increasing number of participating sites.
FeatureCloud is a no-code platform for federated learning apps having the potential to vastly increase the accessibility of privacy-preserving and distributed data analysis in biomedicine and beyond.

Keywords: bioinformatics, data mining, federated learning

Acknowledgement: This project has received funding from the European Union's Horizon2020 research and innovation programme under grant agreement No 826078.

Using AI/ML to transform molecular biology databases

Alex Bateman[1]

[1] EMBL-EBI, Wellcome Genome Campus, Hinxton,
Cambridgeshire, CB10 1SD, UK
agb@ebi.ac.uk

We are living through a revolution in AI approaches, which is transforming molecular biology and computational biology. I will discuss how the advent of high accuracy structural models has made a large impact in our ability to completely and accurately classify protein domains. I will also talk about how Deep Learning models such as ProtENN developed by Google Research have expanded our ability to find distant homologues for known protein families. I will argue that these models represent the most significant change in protein classification in three decades. Even more recently we have seen to arrival of Large Language Models such as ChatGPT, which may now enable us to develop high throughput tools for annotating proteins, non-coding RNAs and families, if only we can stop them hallucinating! I will talk about our efforts to harness these models to write accurate and verifiable annotation at scale.

Keywords: AI, protein domains, molecular biology databases

Stereo-seq: Large Field of View-Spatially Resolved Transcriptomics at Nanoscale Resolution

Javier Batista Perez[1]

[1] BGI Research Foundation Latvia, Lidostas Parks, Marupe, Marupes novads, LV-2167, Latvia
javier1@mgi-tech.com

STOmics' Stereo-seq technology is a world-leading unbiased whole transcriptome spatial omics platform that combines nanoscale resolution with the centimeter-level field of view. The Stereo-seq chip comprises billions of probes, each with spatial barcodes, in a patterned array. Biological tissue is sectioned and loaded to the Stereo-seq chip, followed by fixation and permeabilization. Then mRNA molecules from the tissue section hybridize with the barcoded probes, and reverse transcription produces spatially barcoded cDNA. Following library preparation and sequencing to generate spatially resolved Stereo-seq data from the tissue section. Proprietary cloud-based analysis reconstitutes the data and enables visualization of the transcriptomics of the original tissue section in space, empowering further research and assembly of spatial omics atlases.

Keywords: Spatial omics, single-cell unbiased whole transcriptomics, nanoscale resolution, centimeter-level field of view

"What is life?'': Open quantum systems approach

Andrei Khrennikov[1]

[1] Linnaeus University, International Center for Mathematical Modeling in Physics and Cognitive Sciences, Växjö, SE-351 95, Sweden
Andrei.Khrennikov@lnu.se

Recently the quantum formalism and methodology started to be applied to modeling of information processing in biosystems, mainly to the process of decision making and psychological behavior (but some applications to microbiology and genetics are considered as well). Since a living system is fundamentally open (an isolated biosystem is dead), the theory of open quantum systems is the most powerful tool for life-modeling. In this presentation, we turn to the famous Schrödinger book "*What is life?*" and reformulate his speculations in terms of this theory. Schrödinger pointed out to order preservation as one of the main distinguishing features of biosystems. Entropy has the tendency to increase (*Second Law of Thermodynamics* for isolated classical systems and dissipation in open classical and quantum systems). Schrödinger emphasized the ability of biosystems to beat this tendency. We demonstrate that systems processing information in the quantum-like way can preserve the order-structure expressed by the quantum (von Neumann or linear) entropy. We emphasize the role of the special class of quantum dynamics and initial states generating *the camel-like graphs for entropy-evolution* in the process of interaction with a new environment E:
1) entropy (disorder) increasing in the process of adaptation to the specific features of E};
2) entropy decreasing (order increasing) resulting from adaptation;
3) the restoration of order or even its increase for limiting steady state. In the latter case the steady state entropy can be even lower than the entropy of the initial state.
Such quantum entropy dynamics is illustrated by graphs obtained via numerical simulation for quantum master equation. For simplicity of modelling we consider only quantum Markov dynamics. But the real dynamics of biosystems' states is non-Markovean.

Keywords: Open quantum systems, biosystems, order stability, entropy dynamics, quantum master equation, adaptation to environment, camel-like shape of entropy

Exploiting the linear organisation of omics network embedding spaces

Noël Malod-Dognin*[1], Alexandros Xenos[1], Sergio Doria Belenguer[1], and Nataša Pržulj[1,2,3]

[1] Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
[2] Department of Computer Science, University College London, London WC1E 6BT, UK
[3] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
noel.malod@bsc.es

We are increasingly accumulating large-scale biological omics data that describe different aspects of cellular functioning. These datasets are typically modelled and analyzed as networks. To ease the downstream analyses, recent approaches embed the nodes of a network into a low-dimensional space by using a skip-gram neural network (e.g. DeepWalk, LINE and node2vec). These methods are implicitly factorizing a positive pointwise mutual information (PPMI) matrix, which could be explicitly factorized with Non-negative Matrix Tri-Factorization (NMTF). Importantly, in Natural Language Processing (NLP), word embeddings obtained by using similar approaches showed linear algebraic structures, which allows for answering analogy questions by using simple linear vector operations. Thus, we investigate if we can obtain and exploit similar linear embedding spaces for the biological omics networks.

We initiate the use of the PPMI matrices to capture the neighborhood relationship or the structural (topological) similarities of nodes in the network. By embedding the human Protein-Protein Interaction (PPI) network by factorizing its PPMI matrix representations with NMTF, we demonstrate that the embedding vectors of genes having different Gene Ontology (GO) annotations are linearly separated in the PPI embedding space.

Then, in analogy to the embedding vector of a sentence being obtained as the sum (average) of the embedding vectors of its constituent words in NLP, we show that the embedding vectors of biological functions and of protein complexes can be obtained by averaging he embedding vectors of the genes that participate in then, and that these embeddings can be used to predict protein complex memberships and cancer genes.

Finally, we investigate the embeddings of cancer and control tissue specific PPI networks and show that simple subtractions allow for identifying cancer altered biological functions and cancer genes.

Keywords: bioinformatics, molecular omics networks, network data mining, network embedding

Acknowledgement: This project has received funding from the European Research Council (ERC) Consolidator Grant 770827 and the Spanish State Research Agency AEI 10.13039/501100011033 grant number PID2019-105500GB-I00.

Inverting convolutional neural networks for super-resolution identification of regime changes in epidemiological time series

Jose M. G. Vilar[1, 2]

[1] Biofisika Institute (CSIC, UPV/EHU),
University of the Basque Country
(UPV/EHU), P.O. Box 644, 48080 Bilbao, Spain
[2] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
j.vilar@ikerbasque.org

Inferring the timing and amplitude of perturbations in epidemiological systems from their stochastically spread low-resolution outcomes is as relevant as challenging. It is a requirement for current approaches to overcome the need to know the details of the perturbations to proceed with the analyses. However, the general problem of connecting epidemiological curves with the underlying incidence lacks the highly effective methodology present in other inverse problems, such as super-resolution and dehazing from machine vision. I will present an unsupervised physics-informed convolutional neural network approach in reverse to connect death records with an incidence that allows the identification of regime changes at a single-day resolution. Applied to COVID-19 data with proper regularization and model-selection criteria, the approach can identify the implementation and removal of lockdowns and other nonpharmaceutical interventions with ± 0.9-day accuracy over the span of a year.

Keywords: bioinformatics, physics-informed neural networks, epidemiology

Acknowledgement: J.M.G.V. acknowledges support from Ministerio de Ciencia e Innovacion under grants PGC2018-101282-B-I00 and PID2021-128850NB-I00 (MCI/AEI/FEDER, UE).

Can we use biobanks to study infectious diseases?

Andrea Gelemanović[1]

[1] Mediterranean Institute for Life Sciences,
Meštrovićevo šetalište 45, Split, Croatia
andrea.gelemanovic@medils.hr

Understanding the molecular and environmental basis of diseases in order to improve diagnosis and treatment represent a top priority for researchers. Much of the progress occurred following the growth of various omics technologies and the IT progress in developing large electronic databases capable of storing huge amounts of data. Biobanks represents the most valuable resource for personalized medicine as these are the large collection of various patient samples with well-annotated clinical data which strive to identify possible links between genetic predisposition and disease. A significant step forward are biobanks that are linked to the electronic health records of each participant enabling up-to-date source of relevant medical information and those "deeply phenotyped" for various other omics data, such as microbiome, epigenome, transcriptome, metabolome and proteome.
Since infectious diseases still represent a huge threat to global human health, and host genetic factors have been implied as determining risk factors for observed variations in disease susceptibility, severity, and outcome, during this lecture we will discuss challenges and opportunities of using biobanks as a potential source to study infectious diseases based on the case example of isolated population-based longitudinal biobank "10,001 Dalmatians". Results of a genome-wide association meta-analyses of 14 different infectious-related phenotypes identified 29 infection-related genetic associations, most belonging to rare variants, all of which have a role in immune response. These findings support the concept that host genetic susceptibility to bacterial and viral infections in adults is polygenic, where common variations have very low explained variance and/or "unfortunate" combinations of numerous rare variants. Expanding our understanding of rare variants may help in the construction of genetic panels which might predict an individual's lifetime vulnerability to major infectious diseases. Furthermore, longitudinal biobanks are a valuable source of data for discovering host genetic variations involved in infectious disease susceptibility and severity. Because infectious diseases continue to exert selective pressure on our genomes, a global network of biobanks with access to genetic and environmental data is required to further explain complicated mechanisms underlying host-pathogen interactions and infectious disease vulnerability.

Keywords: biobanks, "10,001 Dalmatians", genome-wide association studies, rare variants, infectious diseases

An integrated platform for genome assembly, comparative genomics and management of genomic variation databases

Jorge Duitama[1]

[1] Systems and Computing Engineering Department.
Universidad de los Andes, Cra 1 Este 19 A 40, Bogotá, Colombia
ja.duitama@uniandes.edu.co

The use of long read DNA sequencing technologies is producing an explosion of high-quality de-novo genome assemblies. The availability of these genomes represents a major step forward for evolution, population genomics, epidemiology, among other applications. A major bottleneck for many research groups continues to be the availability of tools to build and analyze the large datasets of genomes that can be produced using these technologies. In this talk, I summarize the functionalities developed by my research group in the version four of the Next Generation Sequencing Experience Platform (NGSEP) to perform a comprehensive analysis of long and short DNA sequencing reads. First, we designed new algorithms for assembly of haploid and diploid samples from long DNA sequencing reads. A minimizers table is constructed from the reads , using K-mer hash codes calculated from rankings relative to the mode of the k-mer counts distribution. Statistics collected during this process are used as features to build layout paths. For diploid samples, we integrated a reimplementation of the ReFHap algorithm to perform molecular phasing. Benchmark experiments using PacBio HiFi and Nanopore sequencing data for different species show that our solution has competitive contiguity and efficiency, as well as superior accuracy in some cases, compared to other currently used software. We also developed a functionality to perform ortholog identification and gene-based alignment of assembled genomes. Proteomes for each genome are extracted and homology relationships are efficiently predicted building indexes of aminoacid sequences by k-mer ocurrance. Then, genes are clustered in orthogroups based on the topology of the graph induced by the predicted relationships. Gene presence/absence matrices are derived from these orthogroups. If genome assemblies are provided as input, synteny relationships are identified for each pair of genomes. We also implemented algorithms to perform alignment of short and long reads to a reference genome. Based on aligned long reads, we improved the classical variants detector to detect long structural variants. Adding up these developments, NGSEP is a comprehensive tool to perform de-novo and reference-based analysis of DNA sequencing reads in a wide variety of experimental settings to solve different research goals.

Keywords: bioinformatics, algorithms, DNA sequencing, software, genome assembly

Acknowledgement: This work was supported by the Colombian Ministry of Sciences research fund "Patrimonio Autónomo Fondo Nacional de Financiamiento Para la Ciencia, la Tecnología Y la Innovación Francisco José de Caldas" through the grant with contract number 80740-441-2020, awarded to J Duitama. We also wish to acknowledge the support of the IT Services Department and ExaCore-IT Core-facility of the Vice Presidency for Research & Creation at the Universidad de Los Andes that allow us to perform the computational analysis.

Bioinformatics and evolution of non-model organisms

Mikhail S. Gelfand[1,2]

[1]Skolkovo Institute of Science and Technology, Moscow, Russia
[2]A.A.Kharkevich Institute for Information Transmission Problems, Moscow, Russia
mikhail.gelfand@gmail.com

The textbooks are written based on *Escherichia coli*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and mouse biology, with some contribution of less popular model species. However, interesting biology and evolution also happens elsewhere, and I'll tell three such stories that have an evolutionary aspect in common: positive selection on mRNA editing in octopuses and their relatives, tetraplet codons in some ciliate infusoria and their seemingly neutral evolution, and (time permitting) recapitulation of the embryonic transcriptional program in insect pupae.

Keywords: comparative genomics, molecular evolution, coleoids, mRNA editing, Euplotes, frameshfting, genetic code, holometabolous insects, transcriptome, pupal development

Acknowledgement: This is joint work with Mikhail Moldovan, Sofya Gaydukova, Aleksandra Ozerova, Pavel Baranov partially supported by the RSF under grant 23-14-00136 and the RFBR under grant 20-54-14005.

Computational bioengineering for heart disease

Nenad Filipovic*[1], Themis Exarchos[2], and Djordje Jakovljevic[3]

[1] Faculty of Engineering, University of Kragujevac, 34000 Kragujevac, Serbia
[2] Deparment of Informatics, Ionian University, Corfu, Greece
[3] Centre for Health and Life Sciences, Coventry University, United Kingdom
fica@kg.ac.rs

In silico clinical trials are a new paradigm for development of a new drug and medical device. SILICOFCM project is multiscale modeling of familial cardiomyopathy which considers a comprehensive list of patient specific features as genetic, biological, pharmacologic, clinical, imaging and cellular aspects.
The 3D deformable-body represents the left and right ventricle of the heart. Blood flow is modeled during the filling phase by applying the fluid-solid interaction method. The ventricle wall is modeled by 3D brick 8-node solid elements, with fibers that have three-dimensional direction. The Navier-Stokes equations are solved using the ALE formulation for fluid with large displacements of the boundary. The ventricle wall model is simulated by the muscle material model. Muscle fiber orientation is defined by direction vector in 3D prescribed through input data. The outlet blood pressure is used as the boundary condition. At the same time, the wall muscle fibers are activated according to the activation function taken from specific patient measurements.
Computational Platform for Multiscale Modelling in biomedical engineering is results of SGABU project that is served as an educational tool for students and researchers. The platform integrates already developed solutions and various datasets related to cancer, cardiovascular, bone disorders and tissue engineering into one multiscale platform. This will enable further validation and parameterization of models, creation of environment for future trends, e.g. in silico clinical trials, virtual surgery, development of prediction models. InSilc project is devoted to in silico mechanical stent testing within ISO 25539 standards and in silico stent deployment for metallic and biodegradable material.
In-silico projects will connect basic experimental research with clinical study and bioinformatics, data mining and image processing tools using very advanced computer models for drug, stent and patient database in order to reduce animal and clinical studies.

Keywords: bioinformatics, *in silico* clinical trials, data mining, cardiovascular disease

Acknowledgement: This paper is supported by the projects that have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952603 (SGABU project). This paper reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

A Similarity-based Normative Framework for Bio-plausible Neural Nets

Anirvan Sengupta[1,2]

[1] Flatiron Institute, 162 5th Ave, New York, NY 10010, USA
[2] Department of Physics and Astronomy, Rutgers University,
136 Frelinghuysen Rd, Piscataway, NJ 08854, USA
anirvans.physics@gmail.com

In the last decade, Artificial Neural Nets (ANNs), rebranded as Deep Learning, have revolutionized the field of Artificial Intelligence. While these neural nets have their origin in analogy with the neural networks in the brain, in many ways they are trained in ways that are very different from how real neurons learn. For example, to date there is no satisfactory biologically plausible mechanism for backpropagation, the workhorse for training ANNs.
Motivated by this gap, we have looked at alternative normative approaches to neural networks that could give rise to more plausible learning rules. One such approach, which works rather well for representation learning problems, is based on similarity matching or kernel alignment. In this approach, one demands that similar sensory inputs produce similar neural activities. From this rather limited constraint, one can give rise to interesting neural networks performing many common unsupervised learning tasks. I will illustrate, in particular, the case of representing continuous manifolds like spatial information. Here , this approach produces representations very much like place cells in the hippocampus. Consequences of our theory and its relations to some experiments would be discussed. Time permitting, I would touch upon the role of similarity matching in current work in ANNs as well.

Keywords: neural network, brain representation learning

Acknowledgement: I acknowledge long and fruitful collaborations with Yanis Bahroun, Dmitri Chklovskii, Alexander Genkin, Cengiz Pehlevan, Shagesh Shridharan and Mariano Tepper that have informed my view. This work was partly supported by a grant (SF 626323) to the author from the Simons Foundation.

Complexity driven evolution of Alternative splicing

Vladimir Babenko*[1] and Timophey Boltunov[1]

[1] Institute of Cytology and Genetics, Novosibirisk, 640090, Russia
bob@bionet.nsc.ru

Based on the animal model of agonistic interactions, we observed co-varied (linked) alternative exons (LEs) in the genes with alternative splicing phenotype in brain. As a result, we have found 263 positively co-varied pairs, and 26 pairs with negative co-variation. To ascertain the data consistency, we employed three organisms cross-validation: human, mouse and rat with available hippocampus brain region SRA repositories, which supported the co-varied effect of the corresponding exons.

From 142 genes with LE events the maximum LE pairs were observed in insulin – related *Sorbs1* (Sorbin And SH3 Domain Containing 1; 18 LE AS events), and synaptic *Nrcam* (12 LE events). 104 genes maintain only 1 LE pair and 36 genes maintain 2-7 LE pairs. Notably there is a mode at 3 LE pairs per gene (14 genes) in genes vs LE events distribution. GO analysis reveals that the majority of genes maintaining LE events have belong to the synaptic genes, RNA-splicing machinery, and chromatin remodeling.

The 'complexity' (entropic) measure of gene is calculated as $-\sum_{i=1,n}\psi \log 2(\psi)$,

where (Ψ) psi is a percent inclusion rate of a particular AS exon, *n* – number of AS exons in the gene. It is evident that linked AS exons decrease gene complexity rate [3], allowing coordinated splicing in high splicing dynamics rate genes, such as synaptic, RNA processing, chromatin remodeling genes. Herein we speculate if LE AS events are of evolutionary advantage for the high splicing turnover genes working in homeostasis equilibrium.

Next step of the work is to elucidate features providing the linking phenomenon, including mRNA secondary structure, the splicing factor binding sites within and around the corresponding exons.

We will present the results on the issue featuring some complex interactions between exons.

Keywords: alternative splicing, entropy, evolution

Acknowledgement: The study was supported by the Russian Science Foundation (grant no. 19-15-00026).

Pangenomic Alignment: Strings plus Graphs

Travis Gagie[1]

[1] Faculty of Computer Science, Dalhousie University,
6050 University Avenue, Halifax, Nova Scotia, B3H 1W5, Canada
travis.gagie@dal.ca

The use of only one or a few reference genomes for DNA alignment is known to bias research results and medical diagnoses, but aligning against many reference genomes has been problematic. If we represent such a pangenomic reference as a set of strings, then each seed we find in a DNA read may occur in many of the genomes, so even reporting all those occurrences can be slow, and extending and chaining seeds can be infeasible. On the other hand, if we represent them as a graph then --- even apart from the significant technical challenges of indexing graphs --- we may find many chimeric matches. The more of humanity's genetic diversity we try to represent in the graph, the fuzzier it becomes, and the greater the probability of spurious results.

Most research on pangenomic alignment uses either a string representation or a graph representation, but not both. In this talk we first describe how a tool called MONI indexes a pangenomic reference as a set of strings in small space such that later, for each maximal exact match in a given read, we can quickly find that match's length, the position of one of its occurrences in the set of strings, and the lexicographic rank of the suffix starting with that occurrence. We then describe how a tool called MARIA will, when fully implemented, store a pangenomic reference as a graph in small space such that, given MONI's output about a maximal exact match, we can quickly report all the non-chimeric occurrences of that match in the graph.

Combining MONI and MARIA will give us the advantages of working with both strings and graphs: we index the set of reference genomes, the whole set of reference genomes, and nothing but the set of reference genomes, but for each maximal exact match we output relatively few occurrences in the graph, which are easy to use later in a pipeline.

Keywords: pangenomic alignment, reference genomes, data structures, indexing

Acknowledgement: This talk covers results obtained in collaboration with many other researchers, in particular Christina Boucher and Marco Oliva at the University of Florida, Ben Langmead at Johns Hopkins University and Massimiliano Rossi at Illumina, for MONI; and Andrej Baláž, Adrián Goga and Alessia Petescia at Comenius University, Simon Heumos at the University of Tübingen and Jouni at the UCSC Genomics Institute, for MARIA. The author was funded by NSERC grant RGPIN-07185-2020, NSF/BIO grant DBI-2029552 to Christina Boucher, and NIH/NHGRI grant R01HG011392 to Ben Langmead.

Circular Codes in the Genetic Information

Elena Fimmel[1] and Lutz Strüngmann[1]

[1] Mannheim University of Applied Sciences,
Paul-Wittsack Str. 10, 68163 Mannheim, Germany
l.struengmann@hs-mannheim.de
e.fimmel@hs-mannheim.de

Codes are the sets of words over arbitrary alphabets with the property of unique decipherability.
Circular codes are a special class of codes. They are the sets of words with the property of unique recognition of the reading frame for any sequence composed of them and written on a circle. They were introduced by Golomb and Gordon in the 60s under the name of codes with bounded synchronization delay, because they have a strong property of synchronization. For this reason, they play an important role in problems of error correction.
In the middle 90's such a circular code *X* was identified in the genes of bacteria, eukaryotes, plasmids, and viruses by a comprehensive statistical investigation. The code *X* contained the 20 trinucleotides that appeared to be the codons that had the highest preference for the correct reading frame compared to frames 1 and 2. Since then intensive research on circular codes in the genetic information and their potential role in maintaining the correct reading frame during the translation process in the ribosome has been done by various authors. In particular, *X*-motifs were identified in (i) genes "universally" (ii) tRNAs of prokaryotes and eukaryotes; (iii) rRNAs of prokaryotes (16S) and eukaryotes (18S), in particular in the ribosome decoding center where the universally conserved nucleotides G530, A1492, and A1493 are included in the *X*-motif; and (iv) genomes (non-coding regions of eukaryotes). Circular codes have a highly complex structure and the ones found in genes possess additional properties like e.g. self-complementarity that reflect their biological nature.
In our talk we give a short introduction to the theory of circular codes and an overview on the methods from mathematics, statistics and bioinformatics to explore their properties and their biological role. Finally, a possible model of the evolution of the genetic code from the perspective of circular code theory is presented.

Keywords: Circular Codes, Genetic Code, Frame-Shift, Translation

Some Applications of Graph-Based Machine Learning Methods on Biological Data

Mladen Nikolić[1]

[1] Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade, Serbia
mladen.nikolic@matf.bg.ac.rs

Machine learning has made considerable contributions to various fields, most notably by providing methods for predictive modeling and data analysis. Usually, different kinds of data are best modeled by specialized machine learning models, tailored to account for the specifics of the data at hand. Graphs are an expressive data representation most suited for representing relationships between objects. The relationships can be interactions, hierarchies, similarities, or others. Such structures can be found in different kinds of data, including biological ones. Luckily, machine learning toolbox abounds with methods suitable for handling these kinds of data and we consider several applications of such graph-based machine learning methods on biological data. First we discuss tree-like hierarchies over the target variable values and the ways to account for such hierarchies in learning. We consider enzyme classification as a suitable application. Then we discuss hierarchies over the target variable values corresponding to directed acyclic graphs and graph neural network as a suitable model for this kind of data. We consider protein function classification as a suitable application. Finally, we discuss construction of similarity graphs over tabular instances, based on autoencoders and graph representation learning ideas. We consider the application of such techniques to the exploratory analysis of biological data related to expression of schizophrenia.

Keywords: machine learning, graphs, biological data

Acknowledgement: I would like to thank my coauthors and collaborators: Jovana Kovačević, Petar Veličković, Stefan Spalević, Nevena Ćirić, Predrag Janjić, and Stefan Kapunac

From multifunctionality to polypathogenicity with intrinsic disorder

Vladimir N. Uversky[1]

[1] Department of Molecular Medicine and USF Health Byrd Alzheimer's Center and Research Institute, University of South Florida, Tampa, Florida 33612, USA
vuversky@usf.edu

Intrinsically disordered proteins (IDPs) lack stable tertiary and/or secondary structure under physiological conditions *in vitro*. IDPs are characterized by an astonishing multi-level spatiotemporal heterogeneity, with their mosaic structure representing a complex combination of foldons, inducible foldons, morphing inducible foldons, non-foldons, semi-foldons, and unfoldons.

IDPs are highly abundant in nature and have functional repertoire that is very broad and complements functions of ordered proteins. Often, IDPs are involved in regulation, signaling and control pathways, commonly acting as hubs in protein-protein interaction networks. Intrinsic disorder is an important constituent of the proteoform concept, representing one of the important means of functional diversification of the proteinaceous products of a gene. Functions of IDPs may arise from specific disordered forms, from inter-conversion of disordered forms, or from order ↔ disorder transitions. The choice between these conformations is determined by the peculiarities of the protein environment, and many IDPs possess an exceptional ability to differently fold in a template-dependent manner. As a result, many IDPs are capable of conducting multiple functions, with such multifunctionality being linked to their spatiotemporal heterogeneity. Therefore, a correlation between protein structure and function represents a "protein structure–function continuum", where a given protein exists as a dynamic conformational ensemble containing multiple proteoforms characterized by diverse structural features and miscellaneous functions.

IDPs are tightly controlled in the norm by various genetic and non-genetic mechanisms. Alteration in regulation of this disordered regulators are often detrimental to a cell, and many IDPs are associated with a variety of human diseases, such as cancer, cardiovascular disease, amyloidoses, neurodegenerative diseases, diabetes and others. Furthermore, many IDPs are multipathogenic, being associated with the origination and development of a number of different diseases. Therefore, there is a though-provoking interconnection between intrinsic disorder, cell signaling, and human diseases, with polypathogenicity of the involved proteins being linked to their structural plasticity and multifunctionality.

Keywords: intrinsically disordered protein, multifunctionality, polypathogenicity, structure-function continuum

Exploring the impact of rare Copy Number Variants on miRNA genes in CAKUT: Insights from integrated bioinformatic analysis and experimental validation

Ivan Jovanović[1]

[1] Institute of nuclear sciences ''Vinča'', National institite of the Republic of Serbia, Laboratory for radiobiology and molecular genetics, Mike Petrovića Alasa 12-14, 11001 Vinča, Beograd, Srbija
ivanj@vin.bg.ac.rs

Rare copy number variants (CNVs) play a significant role in CAKUT development. However, the specific genetic drivers in certain CNVs associated with CAKUT remain unknown. To explore the genetic elements within CAKUT-associated CNVs, beyond the protein-coding genes, we leveraged the recently described comprehensive CNV landscape of CAKUT. MicroRNAs (miRNAs) are intriguing regulators of genomic networks and have the potential to be involved in CAKUT. Hereby, a pipeline for comprehensive analysis of miRNA genes affected by known, rare CNVs associated with CAKUT will be presented. The procedure is consisted of collection and synchronization of CNV regions specified in different hg assemblies with the hg19 assembly, mapping of the miRNA precursors, identification of the most frequently affected miRNAs and miRNA families by rare CNVs, bioinformatic interpretation of the top-rated miRNAs and prioritisation of key miRNAs for functional validation. Additionally, a method for estimation of the overall burden of rare CNVs on miRNA genes in CAKUT will be discussed. Remarkably, it was found that 80% of CAKUT patients with underlying rare CNV had at least one miRNA gene overlapping the identified CNV. Network analysis of the most frequently affected miRNAs has revealed the dominant regulation of the two miRNAs, hsa-miR-484 and hsa-miR-185-5p. Additionally, miR-548 family members have shown substantial enrichment in rare CNVs in CAKUT. The in vitro model which depicts the heterozygous deletion of the MIR484 has confirmed the study concept implying that rare CNVs affect the corresponding miRNA expression and subsequently dysregulatres miRNA target genes. The translational capacity of miRNA to be employed in therapeutic approaches is nowadays increasingly investigated. Therefore, the untangling of the mechanisms affected by dysregulated miRNAs could serve for future extension of genetic testing and the development of novel miRNA targeting strategies in CAKUT.

Keywords: CAKUT, CNV, miRNA.

Acknowledgement: This work was funded by the Science Fund of the Republic of Serbia, PROMIS, #6066923, miFaDriCa, and the Serbian Ministry of Education, Science, and Technological Development.

Persistence of plasmids targeted by CRISPR interference in bacterial populations

Konstantin Severinov[1]

[1] Waksman Institute for microbiology, Rutgers,
The State University of New Jersey, Piscataway, NJ 08854
severik@waksman.rutgers.edu

CRISPR-Cas systems provide prokaryotes with an RNA-guided defense against foreign mobile genetic elements (MGEs) such as plasmids and viruses. A common mechanism by which MGEs avoid interference by CRISPR consists of acquisition of escape mutations in regions targeted by CRISPR. Here, using microbiological, live microscopy and microfluidics analyses we demonstrate that plasmids can persist for multiple generations in some Escherichia coli cell lineages at conditions of continuous targeting by the type I-E CRISPR-Cas system. We used mathematical modeling to show how plasmid persistence in a subpopulation of cells mounting CRISPR interference is achieved due to the stochastic nature of CRISPR interference and plasmid replication events. We hypothesize that the observed complex dynamics provides bacterial populations with long-term benefits due to continuous maintenance of mobile genetic elements in some cells, which leads to diversification of phenotypes in the entire community and allows rapid changes in the population structure to meet the demands of a changing environment.

Keywords: CRISPR-Cas, plasmid maintenance, plasmid persistence, mathematical modeling

Omics Data Fusion for Understanding Molecular Complexity Enabling Precision Medicine

Nataša Pržulj[1,2,3]

[1] Barcelona Supercomputing Center (BSC), Barcelona 08034, Spain
[2] Department of Computer Science, University College London, London WC1E 6BT, UK
[3] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain

We are flooded by increasing volumes of heterogeneous, interconnected, systems-level, molecular (multi-omic) data. They provide complementary information about cells, tissues and diseases. We need to utilize them to better stratify patients into risk groups, discover new biomarkers, and repurpose known and discover new drugs to personalize medical treatment. This is nontrivial, because of computational intractability of many underlying problems, necessitating the development of algorithms for finding approximate solutions (heuristics).

We develop a versatile data fusion (integration) machine learning (ML) framework to address key challenges in precision medicine from these data: better stratification of patients, prediction of biomarkers, and re-purposing of approved drugs to particular patient groups, applied to cancer, Covid-19, rare thrombophilia and Parkinson's Disease. Our new methods stem from graph-regularized non-negative matrix tri-factorization (NMTF), a machine learning technique for dimensionality reduction, inference and co-clustering of heterogeneous datasets, coupled with novel network science algorithms. We utilize our new framework to develop methodologies for improving the understanding the molecular organization and disease from the omics data embedding space.

An agnostic analysis of the human AlphaFold2 proteome using local protein conformations

Alexandre G. de Brevern[1]

[1] DSIMB Bioinformatics Team, INSERM UMR_S 1134, BIGR, Université Paris Cité and Université de la Réunion, 75014 Paris, France
alexandre.debrevern@univ-paris-diderot.fr

For more than 30 years, different computational approaches have been implemented to propose 3D structural models of proteins from their amino acid sequence. Using deep Learning, AlphaFold 2 obtained particularly remarkable results; some models were within the uncertainties of the experimental resolution (Jumper et al., *Nature* 2021). AlphaFold 2 code is freely avalaible and EBI provides structural model databases (Tunyasuvunakool et al., *Nature* 2021), i.e. 98.5% of the human proteome is given. 36% of these models are predicted with atomistic quality.
The human protein models provided by AlphaFold were analyzed using its confidence index (pLDDT score), with classic secondary structure and finer analysis of local protein conformation, e.g. γ-turns, β-turns and bends, β-turn types, PolyProline II (PPII), helix curvatures, β-bulges, and a structural alphabet, namely Protein Blocks (PB).
As expected, the large majority of α-helices are well predicted with high pLDDT scores. However, some points are intriguing and could potentially lead to improvements in the future: (i) PPII helices are too often encountered with a low confidence index. They represent 4-5% of all residues and are important in protein-protein interactions; it could so be an issue to be poorly approximated. (ii) In a very surprising way, while β-turns (turns of 4 residues) are well predicted, 55% of γ-turns (3 residues) have very low pLDDT scores. (iii) Even more strikingly, 94.8% of cis ω angles associated with low pLDDT scores, i.e. AlphaFold is clearly unable to propose proper cis ω angles. (iv) β-sheet occurrence is lower than expected, while PB *d* (i.e. β-sheet core geometry) occurrence is completely in accordance with the expected frequencies. There are so potentially β-sheets that were not founded until the end, which would explain this low frequency (de Brevern, *Biochimie* 2023). AlphaFold 2 had impacted the structural modeling area but works remained (Tourlet et al., *BioMedInformatics* 2023)

Keywords: bioinformatics, deep learning, computer science, protein structure, secondary structures.

Uncovering resistance to microtubule targeting drugs

Mattia Pavani[1], Elena Chiroli[1], Paolo Bonaiuti, and Andrea Ciliberto*[1,2]

[1]IFOM, The AIRC Institute of Molecular Oncology,
via Adamello 16, 20139, Milan, Italy
[2]Pazmany Peter Catholic University, Faculty of Information Technology and Bionics, 1083, Budapest, Hungary
andrea.ciliberto@ifom.eu

Drugs that alter microtubule dynamics have been used for decades for treating different types of cancers. Such drugs arrest cell cycle progression in mitosis, and induce apoptosis. However, a fraction of cells manages to survive, escapes from the arrest and resumes proliferation. Understanding the strategies of these cells is important to uncover the early stages of emergence of resistance. To this aim, we performed laboratory evolution experiments in yeast and in mammalian cells when microtubule dynamics is impaired. Our results show that cells follow reproducible strategies to escape the effect of the drug. Via mutations, aneuplouidy and non genetics mechanisms they recover microtubule functionality and decrease the propensity to die.

Keywords: cell division, resistance, microtubules, molecular networks

Acknowledgement: The studies we will present were financed by AIRC, the italian association for cancer research and by the hungarian national research, development and innovation office.

Computational tools and repositories for precision therapeutics in the post-genomic era

George P. Patrinos[1,2,3]

[1] University of Patras School of Health Sciences, Department of Pharmacy, Patras, Gtreece
[2] United Arab Emirates University, College of Medicine and Health Sciences, Department of Genetics and Genomics, Al-Ain, UAE
[3] Erasmus MC, Faculty of Medicine and Health Sciences, Department of Pathology, Clinical Bioinformatics Unit, Rotterdam, the Netherlands.
gpatrinos@upatras.gr

In the post-genomic era, the rapid evolution of high-throughput genotyping technologies and the increased pace of production of genetic research data, are continually prompting the development of appropriate informatics tools, systems and databases as we attempt to cope with the flood of incoming genetic information. Alongside new technologies that serve to enhance data connectivity, emerging information systems should contribute to the creation of a powerful knowledge environment for genotype-to-phenotype information in the context of translational medicine. In the area of pharmacogenomics and personalized medicine, it has become evident that database applications providing important information on the occurrence and consequences of gene variants involved in pharmacokinetics, pharmacodynamics, drug efficacy and drug toxicity, will become an integral tool for researchers and medical practitioners alike. At the same time, two fundamental issues are inextricably linked to current developments, namely data sharing and data protection. In this lecture, the impact of high throughput and next generation sequencing technology and its impact on pharmacogenomics research and clinical implementation of genomic medicine will be addressed. In addition, advances and challenges in the field of pharmacogenomics information systems will be discussed, which in turn prompted the development of an integrated electronic 'pharmacogenomics assistant'. The system is designed to provide personalized drug recommendations based on linked genotype-to-phenotype pharmacogenomics data, as well as to support biomedical researchers in the identification of pharmacogenomic related gene variants.

Development of hybrid and optimized deep learning classifiers for speech recognition in tracheostomy patients: a case study

Themis Exarchos[1]

[1] Department of Informatics, Ionian University,
Ioannou Theotoki 72, Corfu, Greece
themis.exarchos@gmail.com

The ability of a machine or program to recognize words spoken aloud and translate them into legible text is known as speech recognition and has gained a lot of attention in the last decade especially in healthcare to promote the quality of care. The problem of speech recognition in patients with tracheostomy has not yet been investigated in the literature. In this work, we propose a hybrid and highly scalable deep learning workflow which utilizes both CNN and RNN architectures across video recordings to identify speech. Dropout rates were also used to avoid overfitting effects. Hyperparameter optimization was applied using the GridSearch method to fine tune the DL workflow on each patient. A case study was applied, where video records were collected from 25 patients in Greece who read specific texts from Greek language, selected by logotherapy experts. A fully automated data processing pipeline was initially applied to extract the video frames based on the provided annotations by the experts (start time, end time per word). Then, we handled the speech recognition problem as a multiclass classification problem, where each word represents a class. Two different types of models were developed; 25 personalized models, which were trained and tested across each individual patient, and a generalized model which was trained and tested on randomly selected instances from all patients. Our results highlight the increased accuracy in terms of reduced word error rate in both the personalized and the generalized hybrid DL models against the conventional DL models.

Keywords: deep learning, speech recognition, tracheostomy

Acknowledgement: This work is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 952603 (SGABU). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

Advancing Genomics with OrthoDB, BUSCO, and the LEM Framework

EV Kriventseva[1], M Manni[1], M Seppey[1], F Tegenfeldt[1], M Berkeley[1], D Kuznetsov[1], EM Zdobnov[1]

[1] Department of Genetic Medicine and Development, University of Geneva Medical School, Swiss Institute of Bioinformatics, rue Michel-Servet 1, 1211 Geneva, Switzerland.
Evgenia.Kriventseva@unige.ch

The rapid growth of genomics data necessitates continuous advancements in bioinformatics tools. This presentation highlights the latest updates to our toolbox, including OrthoDB v11, BUSCO v5, and the LEM benchmarking framework.
OrthoDB (https://www.orthodb.org) is a leading resource for gene orthology and functional annotations across diverse eukaryotes, prokaryotes, and viruses. Orthology facilitates precise bridging of gene function knowledge within the genomics sphere. OrthoDB v11 encompasses over 100 million genes from 18,000 prokaryotes and nearly 2,000 eukaryotes, providing extensive species coverage. The open-source OrthoLoger software (https://orthologer.ezlab.org) allows mapping of novel gene sets to precomputed orthologs, linking them to relevant annotations.
BUSCO (https://busco.ezlab.org) serves as a standard tool for assessing the completeness of genome assemblies, transcriptomes, and predicted gene sets, complementing assembly contiguity measures like N50 values. A spin-off of OrthoDB, BUSCO evaluates the presence and coverage of marker genes, offering an evolutionarily-grounded expectation of gene content completeness. BUSCO v5 now automatically selects the most suitable dataset for evaluation, outperforming the popular CheckM tool. Its efficiency is particularly evident in large eukaryotic genomes, and it is uniquely capable of assessing both eukaryotic and prokaryotic species, making it applicable to metagenome-assembled genomes of unknown origin.
The LEMMI (https://lemmi.ezlab.org) benchmarking framework, now in version 2, facilitates informed software tool selection. This Live Evaluation of Methods (LEM) for Metagenome Investigation uses a container-based approach for continuous benchmarking and effective end-user distribution. The versatile framework can be extended to other procedures, such as gene orthology inference with LEMOrtho (https://lemortho.ezlab.org). The LEM benchmarking approach aims to become a community-driven effort, allowing developers to showcase novel methods and users to access standardized, easy-to-use software. We encourage researchers to apply this framework in their domain and welcome feedback.

Keywords: genomics, genomes, orthologs, genes, continuous benchmarking

Multiomics Integration by Non-Negative Tri-Matrix Factorization Reveals New Target Genes in Parkinson's Disease

Alexander Skupin[1]

[1] Luxembourg Centre for Systems Biomedicine, Belvaux, Luxembourg

Parkinson's disease (PD) is the second most common neurodegenerative disease which is characterized by neuronal loss of dopaminergic neurons (mDA) in the substantia nigra. The underlying complexity of the disease and limited amount of patient material limits current interventions to only symptomatic and no curative treatment despite intensive research. We use patient-derived induced pluripotent stem cells to generate mDAs and investigate disease mechanisms by multiomics characterization including single cell RNA-sequencing and bulk proteomics and metabolomics. For this purpose, we developed an extended Non-Negative TriMatrix Factorization approache that allows to integrate the heterogeneous omics data with knowledge of molecular databases including protein-protein, genetic and metabolic interactions as well as co-expression profiles. Our approach was able to identify already PD-associated but also new druggable candidate genes of PD development.

Prediction of cell types using single-cell mRNA profiles

Vladimir Brusic[1]

[1] University of Nottingham, Ningbo, China

Single cell transcriptomics is a rapidly growing area with an urgent need for new analytical tools to complement and supersede unsupervised clustering. We defined a new method for deriving gene expression profiles from single-cell gene expression matrices. We named these profiles the "single-cell-derived-class" (SCDC) profiles. We developed SCDC profiles for multiple cell types and subtypes of peripheral blood mononuclear cells (PBMC) using the results of single cell transcriptomics (SCT) experiments. SCDC profiles represent characteristic patterns of gene expressions of the types and subtypes of healthy human PBMC. We studied the reproducibility of SCDC profiles, their robustness, and their applications in classifying healthy human PBMC types and subtypes. SCDC profiles are efficient and convenient tools for the analysis of SCT data derived from PBMC samples. These profiles are highly reproducible, even when derived from unrelated studies, provided that the sample processing steps are comparable and the same SCT technology is used. The classification accuracy of SCDC profiles is high. SCDC profiles can be used for supervised classification and the discovery of new subtypes of PBMC.

The complete solution and interpretation algorithms for large field-of-view and high-resolution spatial transcriptomics

Shuangsang Fang[1]

[1] BGI-Research, Belgrade, Serbia

The large field-of-view and high-resolution spatial transcriptomics technology can reveal and answer scientific questions that cannot be discovered or elucidated by low-resolution spatial transcriptomics. Obtaining expression profiles at the single-cell level from high-resolution spatial transcriptomics requires sophisticated data processing and interpretation strategies, including extensive image data processing, transcriptome data processing, integration analysis. At the same time, the introduction of spatial information helps with the annotation of single cells at the tissue level and the study of tissue structure and function, while cell clustering and cell annotation are important foundations for subsequent in-depth analysis. Cell annotation can be divided into clustering and re-annotation based on marker genes and end-to-end cell annotation based on reference datasets. The choice between the two depends on whether markers are easier to obtain or whether reference datasets with consistent data backgrounds are easier to obtain. The algorithm team at BGI Research Institute has conducted extensive algorithm research and development on data interpretation strategies, cell clustering algorithms, and cell annotation algorithms for large field-of-view and high-resolution spatial transcriptomics technology, with the aim of providing comprehensive, efficient, and highly reliable data analysis algorithms, tools and platform support.

To be folded, to be unfolded or to be aggregated with important functions: application of the directed coaggregation mechanism to combat bacterial communities

O.V. Galzitskaya*[1], S.Yu. Grishin[1], A.V. Glyakina[1], M.V. Slizen[1], A.V. Panfilov[1], P.A. Domnin[2,3], A.P. Kochetov[4,5], A.A. Surin[6], S.V. Kravchenko[7], A.K. Surin[5], S.A. Ermolaeva[2]

[1] Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino, Russia
[2] Gamaleya Research Centre of Epidemiology and Microbiology, 123098 Moscow, Russia
[3] Biology Faculty, Lomonosov Moscow State University, 119991 Moscow, Russia
[4] Pushchino State Institute of Natural Sciences, Pushchino, Moscow Region 142290, Russia
[5] The Branch of the Institute of Bioorganic Chemistry, Pushchino, Russia
[6] Faculty of Applied math, MIREA - Russian Technological University, Moscow, 119454 Russia
[7] Institute of Environmental and Agricultural Biology (X-BIO), Tyumen State University, 625003 Tyumen, Russia

ogalzit@vega.protres.ru

One of the reasons for the mortal danger to humans is the ability of pathogenic bacteria to form biofilms. The formation of biofilms is an evolutionarily conservative defense mechanism against adverse conditions. The use of this protection by pathogenic bacteria reduces the effectiveness of the main means of combating them - antibiotics, which complicates the production of new types of drugs. There are two types of antimicrobial agents that are not known antibiotics: nanoparticles and antimicrobial peptides. We demonstrated that peptides synthesized based on the amino acid sequence of proteins and capable of amyloid formation and coaggregation with the whole protein exhibit antimicrobial activity. The ability of peptides to coaggregate with target proteins can help combat biofilm-forming bacterial communities.

We evaluated the antimicrobial effects of ten synthesized hybrid peptides, which were obtained based on the sequences of the S1 ribosomal protein of *P. aeruginosa* and *S. aureus*. It is important that some peptides demonstrated high antimicrobial activity comparable to the antibiotic gentamicin sulfate against pathogenic strains of MRSA, *S. aureus*, and *P. aeruginosa*. These peptides showed no toxicity to eukaryotic cells. Our study demonstrates the promise of hybrid peptides based on the amyloidogenic regions of the S1 ribosomal protein for the development of new antimicrobials against Gram-positive and Gram-negative bacteria resistant to traditional antibiotic.

Keywords: amyloid, coaggregation, antimicrobial peptides

Acknowledgement: This research was funded by the Russian science foundation, Grant Number 18-14-00321.

Oral presentation

CADDIE - An online knowledge base for network-based mechanism exploration and drug repurposing in oncology

Michael Hartung*[1], Elisa Anastasi[2], Zeinab M. Mamdouh[3,4], Cristian Nogales[3], Harald HHW Schmidt[3], Jan Baumbach[1,5], Olga Zolotareva[1,6], and Markus List[6]

[1]Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany
[2]School of Computing, Newcastle University, Newcastle upon Tyne, UK
[3]Department of Pharmacology and Personalised Medicine, Maastricht University, Maastricht, Netherlands
[4]Department of Pharmacology and Toxicology, Faculty of Pharmacy, Zagazig University, Zagazig, Egypt
[5]Computational Biomedicine Lab, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[6]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising, Germany
michael.hartung@uni-hamburg.de

Drug repurposing is the use of previously developed and tested pharmaceutical agents in new application cases and lately often used as a solution to the increasing drug development costs. Cancers are extremely heterogeneous disorders demonstrating a wide variability of drug responses due to diverse subtypes, quickly evolving and acquiring drug resistance. Therefore, the identification of compounds that can effectively combat a specific tumor type is crucial. Drug candidates that are potentially effective against a specific tumor can be chosen based on the set of driver mutations acquired by this tumor. For optimal treatment, it is important to consider targeted anti-cancer therapies and drugs initially developed to treat non-cancerous diseases.
To overcome this hurdle, we present CADDIE (Cancer Driver Drug Interaction Explorer), a web platform to identify oncological drug repurposing candidates. CADDIE's biomedical knowledge base integrates a multitude of gene-gene and drug-gene interaction datasets, detailed anticancer drug information and cancer biology data such as cancer driver genes, mutation frequencies and gene expressions. For the purpose of locating drug targets and candidates for drug repurposing, CADDIE makes network medicine algorithms available to the researchers. It guides the users from the choice of seed genes through the discovery of therapeutic targets or drug candidates. Network medicine also provides indirect strategies that take into account other functionally relevant targets in the gene interaction network since potential cancer driver genes may be inaccessible for direct targeting. We demonstrate the application of CADDIE in different cancer subtypes such as sarcoma and ovarian cancer with a detailed analysis of the found drug targets and chemical compounds. CADDIE is available online at https://exbio.wzw.tum.de/caddie/ and as a python package at https://pypi.org/project/caddiepy/.

Keywords: Drug repurposing, Drug prioritization, Network medicine, Cancer, Network Analysis, Network Medicine

Acknowledgement: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. This project reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the *e:Med* research and funding concept (*grants 01ZX1910D and 01ZX2210D*). JB was partially funded by his VILLUM Young Investigator Grant nr.13154.; Z.M. is funded by a full scholarship [40463/2019] from the Ministry of Higher Education of the Arab Republic of Egypt.

Mapping of Disease Names to Disease Codes based on Natural Language Processing Techniques

Anđelka Zečević*[1], Jovana Kovačević[2], and Radoslav Davidović[3]

[1]Mathematical Institute, Serbian Academy of Sciences and Arts
[2]Faculty of Mathematics, University of Belgrade
[3]Institute of Nuclear Sciences Vinča
andjelkaz@mi.sanu.ac.rs

Information aggregation from various gen, disease, and gen-disease databases such as DisGeNet, COSMIC, HumsaVar, Orphanet, ClinVar, HPO, and Diseases into a unique database would enable researchers to analyze and compare valuable domain findings in a more convenient and systematic way. However, the aggregation poses numerous challenges due to non-uniform information annotation across the databases. In this work, we address the problem of mapping a disease name, when needed, into a standardized disease code (DOID) based on Natural Language Processing text representation techniques. We examine the benefits and limitations of using off-the-shelf embeddings such as Med2vec, and language models such as BioBERT, UmlsBERT, and PubMedBERT in retrieval scenarios with respect to standard full-text search. In addition to qualitative improvements, we elaborate on the technical requirements and computational complexities that come with the embracement of language models and semantic search.

Oral presentation

Zero- and Few-Shot Machine Learning for Named Entity Recognition in Biomedical Texts

Miloš Košprdić*[1], Nikola Prodanović[1], Adela Ljajić[1],
Bojana Bašaragin[1], and Nikola Milošević[1,2]

[1]Institute for Artificial Intelligence Research and Development of Serbia,
Fruškogorska 1, Novi Sad, Serbia
[2]Bayer A.G., Reaserch and Development, Mullerstrasse 173, Berlin, Germany
milos.kosprdic@ivi.ac.rs

Named entity recognition (NER) is an NLP that involves identifying and classifying named entities in text. Token classification is a crucial subtask of NER that assumes assigning labels to individual tokens within a text, indicating the named entity category to which they belong. Fine-tuning large language models (LLMs) on labeled domain datasets has emerged as a powerful technique for improving NER performance. By training a pre-trained LLM such as BERT on domain-specific labeled data, the model learns to recognize named entities specific to that domain with high accuracy. This approach has been applied to a wide range of domains including biomedical and has demonstrated significant improvements in NER accuracy.

Still, data for fine-tuning pre-trained LLMs is large and labeling is a time-consuming and expensive process that requires expert domain knowledge. Also, domains with an open set of classes yield difficulties in traditional machine learning approaches since the number of classes to predict needs to be pre-defined.

Our solution to the two mentioned problems is based on data transformation for factorizing the initial multiple classification problem into a binary one and applying cross-encoder-based BERT architecture for zero- and few-shot learning.

To create our dataset, we transformed six widely used biomedical datasets that contain various biomedical entities such as genes, drugs, diseases, adverse events, chemicals, etc., into a uniform format. This transformation process enabled us to merge the datasets into a single cohesive dataset of 26 different named entity classes.

We then fine-tuned two pre-trained language models: BioBERT and PubMedBERT for the NER task in zero- and few-shot settings. The results of the experiment for 9 classes in zero-shot mode are promising for semantically similar classes and improve significantly after providing only a few supporting examples for almost all classes. The best results were obtained using a fine-tuned PubMedBERT model, with average F1 scores of 35.44%, 50.10%, 69.94%, and 79.51% for zero-shot, one-shot, 10-shot, and 100-shot NER respectively.

Keywords: zero-shot learning, machine learning, deep learning, natural language processing, biomedical named entity recognition

Clustering and classification of SARS-COV-2 isolates using RSCU

S. Malkov*[1], M. Beljanski[1], G. Pavlović Lažetić[1], B. Stojanović[2], M. Maljković[1], A. Veljković[1], S. Kapunac[1], and N. Mitić[1]

[1]Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[2]Mathematical Institute SASA, Knez Mihaila 36, 11000 Belgrade, Serbia
sasa.malkov@matf.bg.ac.rs

The existence of a large number of sequenced SARS-COV-2 isolates provides an opportunity to observe genomic variability in a massive sample. The goal of our research was to use data mining techniques to study possible correlation between codon usage and classification by WHO-labels in a certain period of time.
The material includes 745,533 isolates with 12,236,672 coding sequences (proteins) from NCBI (10.08.2022.). RSCU was used as a measure of codon usage. Samples are associated with WHO-labels (based on Pango_Id) and time intervals. Inconsistency of WHO-labels with periods in which the respective strains were actually present was observed. The isolates with the observed discrepancy were excluded from the sample. Isolates without assigned WHO-labels were also excluded. In addition, individual coding sequences containing ambiguous nucleotide codes were eliminated.
Clustering was performed for each of the 12 common types of coding sequences (proteins), with multiple methods and a different number of clusters. Neural clustering gave the best results. For different protein types, different degrees of RSCU variability are observed. In the case of proteins with a small variation in nucleotide contents, over 95% of the material belongs to a single cluster, while the other clusters are of negligible size. In the case of proteins with more variations, a higher number of pure clusters (by WHO-labels) is obtained, with a small number of heterogeneous clusters (about 10% of the material). In those heterogeneous clusters, there are isolates with different WHO-labels that were present in parallel at some point, as a kind of transitional forms between two strains.
Different classification models were created on the same sample. Models based on protein types with higher diversity between coding sequences are highly accurate (96-100%). Using the classification models, the corresponding WHO-labels were associated with isolates without previously assigned WHO-labels.

Keywords: SARS-COV-2, RSCU, clustering, classification

Oral presentation

The use of Active Machine Learning for Protospacer-Adjacent Motif recovery in Class 2 CRISPR-Cas systems

Bogdan Kirillov*[1,2], Aleksandra Vasileva[3,4], Oleg Fedorov[5], Maxim Panov[6], and Konstantin Severinov[7]

[1]Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld 1, 121205 Moscow, Russia
[2]Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, 34/5 Vavilova Street, 119334 Moscow, Russia
[3]Peter the Great St. Petersburg Polytechnic Universitsy, Politekhnicheskaya St 29, 195251 St. Petersburg, Russia
[4]Institute of Molecular Genetics, Russian Academy of Sciences, Kurchatov square 2, 123182 Moscow, Russia
[5]Research Institute for Systems Biology and Medicine, Department of Mathematical Biology and Bioinformatics, Nauchny Proezd 18, 117246 Moscow, Russia
[6]Artificial Intelligence Cross Center Unit, Technology Innovation Institute, PO Box: 9639, Masdar City, Abu Dhabi, United Arab Emirates
[7]Waksman Institute of Microbiology, Rutgers, State University of New Jersey, Piscataway, NJ 08854, USA
Bogdan.Kirillov@skoltech.ru

The recognition of target DNA sequences during the interference phase of prokaryotic CRISPR-Cas immunity relies on Protospacer-Adjacent Motif (PAM) sequences, specific for each Cas effector. PAM identification is a laborious and time consuming process that requires multiple stages including *in vitro* and *in vivo* cleavage assays followed by Next Generation Sequencing of targets that withstood cleavage. Determining PAM is an essential step of characterisation of any novel Cas9 ortholog and determines the likelihood of its potential use. This study investigates the potential of machine learning to predict PAM sequences for a given Cas9 ortholog based on the results of cleavage experiments and employing an Active Learning process akin to Reinforcement Learning with Human Feedback. Machine learning-facilitated PAM identification would streamline and accelerate existing pipelines for describing novel Cas proteins. We demonstrate that simple models with a small amount of data are sufficient for confident PAM predictions when training is effectively orchestrated.

Keywords: bioinformatics, CRISPR, machine learning

Application of classification algorithms for hip implant surface topographies

Aleksandra Vulović*[1,2], Tijana Geroski[1,2], and Nenad Filipović[1,2]

[1] Faculty of Engineering University of Kragujevac,
Sestre Janjić 6, 34000 Kragujevac, Serbia
[2] Bioengineering Research and Development Center (BioIRC),
Prvoslava Stojanovića 6, 34000 Kragujevac, Serbia
aleksandra.vulovic@kg.ac.rs

Experimental studies have shown that lower shear stress values lead to better femoral bone – hip implant connection. Numerical simulations have provided option to reduce the number of experimental studies through analysis of different hip implant surface topographies. However, this approach takes time as there are different model parameters that should be considered in order to understand how they affect the obtained shear stress values. The use of classification algorithms is an approach that could reduce the time required for simulation by providing information about models with biggest potential. Eleven model parameters related to model and surface topography were considered in combination with four classification algorithms - Support Vector Machines (SVM), K - Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF). The considered parameters were: Number of half-cylinders lengthwise (>0); Number of half-cylinder rows (≥0); Half cylinders added or removed from the surface (0 – removed; 1 - added); Distance between half-cylinders lengthwise (≥0); Distance between half-cylinders widthwise (≥0); Number of different radius values (1 or 2); Radius 1 value (>0); Radius 2 value (≥0); Distance from the edge where loading is located (≥0); Distance from the other edge of the model (≥0); Model includes trabecular bone (0 – not included; 1 - included). The aim was to apply previously mentioned algorithms to obtain information if the maximum shear stress value was above or below user-defined threshold. The obtained results show that this approach can be useful to obtain preliminary information about models that should be numerically analyzed.

Keywords: classification, finite element analysis, hip implant, surface topographies

Acknowledgement: This research is supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 952603 - SGABU. This article reflects only the author's The Commission is not responsible for any use that may be made of the information it contains. Authors also acknowledge the funding by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract number [451-03-47/2023-01/200107 (Faculty of Engineering, University of Kragujevac)].

Oral presentation

Computational Modelling of Drug Effects on Cardiomyopathy and Analysis of Myocardial Work

Smiljana Tomasevic[1,2]*, Miljan Milosevic[2,3], Bogdan Milicevic[1,2], Vladimir Simic[2,3], Momcilo Prodanovic[2,3,4], Srboljub M. Mijailovich[4,5], 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, 34000 Kragujevac, Serbia
[3] Institute for Information Technologies, University of Kragujevac, Jovana Cvijica bb, 34000 Kragujevac, Serbia
[4] FilamenTech, Inc., Newton, MA 02458, USA
[5] BioCAT, Department of Biology, Illinois Institute of Technology, Chicago, IL 60616, USA
smiljana@kg.ac.rs

Analysis of myocardial work is essential in determination of left ventricle ejection fraction (LVEF) and non-invasive assessment of different types of cardiomyopathies. Two major classifications of cardiomyopathy are: dilated (DCM) and hypertrophic (HCM) cardiomyopathy. Although there are clinical improvements in cardiomyopathy risk assessment, patients are still under high risk of severe events. Computational modeling of and computer-aided drug design can significantly advance the understanding of cardiac muscle activity in DCM and HCM cardiomyopathies, speed up the drug discovery and reduce the risk of severe events, aiming to improve the treatment of cardiomyopathy.
The main advantage and novelty of presented study are coupled macro and micro simulations into the integrated Fluid Solid Interaction (FSI) system and its application for examination of heart behavior and drug interactions. In contrary to detailed and patient-specific models where FSI analyses are very time-consuming, our models are parametric and based on dimensions of specific LV components. FSI algorithm within the PAK software is used for modeling the LV with nonlinear material model, together with stretches integration along muscle fibers. The methods are integrated within the SILICOFCM platform, and aim to propose an advanced approach for the assessment of work indices and biomechanical characteristics of cardiomyopathies and drugs effects, based on computational modelling.
In this study, simulations of the effect of drugs on improving performance of DCM LV parametric model include the drugs that affect calcium transients (Dygoxin) and changes in kinetic parameters (2-deoxy adenosine triphosphate - dATP). Myocardial word is presented through changes of pressures and volumes (P-V diagrams) for DCM LV model at basic condition (without administered drug) and with using Dygoxin and dATP. Due to increased LV size, the P-V loop for the DCM model without administered drug is shifted toward lower ventricular pressure and lager ventricular volume, with LVEF = 56.83%. Effects of drugs on DCM show an increase in ventricular peak pressures and LVEFs, while the P-V loops are shifted toward decreased volumes, corresponding to healthy hearts.
Computational modeling and drug design approaches can speed up the drug discovery and significantly reduce expenses aiming to improve the treatment of cardiomyopathy.

Keywords: Computational modelling, Myocardial work, Dilated cardiomyopathy, Drug effects

Acknowledgement: This work is supported by the European Union's Horizon 2020 research and innovation pro-grammes SILICOFCM (Grant agreement 777204) and SGABU (Grant agreement 952603). The Commission is not responsible for any use that may be made of the information it contains. The research was also funded by Serbian Ministry of Education, Science, and Technological Development, grants [451-03-47/2023-01/200378 (Institute for Information Technologies, University of Kragujevac)] and [451-03-47/2023-01/200107 (Faculty of Engineering, University of Kragujevac)].

Echocardiography-based Left Ventricle Cardiac Hypertrophy Simulations

Bogdan Milićević*[1,2], Miljan Milošević[2,3,4], Vladimir Simić[2,3], Danijela Trifunović[5], Goran Stanković[5,6], Nenad Filipović[1,2], and Miloš Kojić[2,6,7]

[1] Faculty of Engineering, University of Kragujevac, Kragujevac 34000, Serbia
[2] Bioengineering Research and Development Center (BioIRC), Kragujevac 34000, Serbia
[3] Institute for Information Technologies, University of Kragujevac, Kragujevac 34000, Serbia.
[4] Belgrade Metropolitan University, Belgrade 11000, Serbia.
[5] Cardiology Department, University Clinical Center of Serbia, Visegradska 26, 11000 Belgrade, Serbia
[6] Serbian Academy of Sciences and Arts, Belgrade 11000, Serbia.
[7] Houston Methodist Research Institute, Houston TX 77030, USA.
bogdan.milicevic@uni.kg.ac.rs

Clinical scenarios can be evaluated using numerical modeling of the cardiac cycle prior to experimental or clinical application. Changes in wall thickness, displacement fields, and general cardiac function are all affected by hypertrophy. In our study, we calculated the effects of eccentric and concentric hypertrophy and monitored changes in ventricular thickness and shape. Concentric hypertrophy results in thicker walls, while eccentric hypertrophy results in thinner walls. Passive stresses were calculated using recently established material modals based on Holzapfel's work. Our modeling approach is based on composite shell finite elements, allowing easier and more efficient modeling compared to traditional 3D finite elements. A left ventricular model was constructed using echocardiographic images. Our modeling technology is based on accurate patient-specific geometries and realistic constitutive curves, so it can be used as the basis for real-world applications. Our model can be used to test medical hypotheses about the development of hypertrophy in healthy and diseased hearts under the influence of different conditions and factors.

Keywords: composite shell finite elements, echocardiography, left ventricle, cardiac hypertrophy

Acknowledgement: This research was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 952603 (http://sgabu.eu/). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains. Research was also supported by the SILICOFCM project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777204. This article reflects only the authors' views. The European Commission is not responsible for any use that may be made of the information the article contains. The research was also funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia, contract numbers [451-03-68/2022-14/200107 (Faculty of Engineering, University of Kragujevac) and 451-03-68/2022-14/200378 (Institute for Information Technologies Kragujevac, University of Kragujevac)].

Oral presentation

Decoding Cystic Fibrosis Phenotype

Aleksandra Divac Rankov*[1], Dušan Ušjak[1],
Martina Mia Mitić[1], and Jelena Kusic Tisma[1]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia
aleksandradivac@imgge.bg.ac.rs

Cystic fibrosis (CF) is a monogenic autosomal recessive disease caused by mutations in transmembrane conductance regulator (CFTR) gene. The golden standard for the diagnosis of CF is sweat chloride testing (>60 mmol/L) together with the identification of two CF-causing variants of CFTR gene. Nevertheless, about 0.01% of patients with elevated sweat chloride and high clinical suspicion of CF do not carry any CF-causing variants.
Here we present analysis of whole exome sequencing (WES) results for two patients with elevated sweat chloride levels and clinical presentation of CF in whom no CF-causing mutations were detected after CFTR gene whole coding region sequencing, and large insertion/deletion testing.
Genomic DNA was extracted from whole blood, subjected to library preparation using DNA nanoball technology from BGI and sequenced on DNBSEQ-G400 (MGI). Produced fastq files were mapped to hg38 reference genome using BWA/SAM tools. VCF files were generated using GATK (BaseRecalibrator, HaplotypeCaller) and annotated with InterVar and AnnoVar tools. Filtering of detected variants for disease relevance was done using the following criteria: QC Filter, GnomAD Allele Frequency, Functional consequences and phenotype-genotype relationship.
In both patients, similar number of variants predicted to impair protein function were detected (27 and 25). In two genes (CACNA1H and MUC5B) missense type variants were found in both patients and loss of function variants were found in 7 and 11 genes, respectively. Functional assessment of selected variants is underway.
Bioinformatics analyses are a valuable tool enabling identification of underlining genetic bases of disease phenotype, important in the context of optimal patient management and targeted therapies.

Keywords: whole exome sequencing (WES), cystic fibrosis, variant assessment

Acknowledgement: This work was supported by IMGGE work program for 2023, Ministry of Education, Science and Technological Development of the Republic of Serbia, 451-03-47/2023-01/200042.

Single cell 3' transcriptome profiling

Nevena Milivojević*[1], Uršula Prosenc Zmrzljak[2], Biljana Ljujić[3], Valentina Đorđević[4], Marina Gazdić Janković[3], Marko Živanović[1,5], Feđa Puač[6], Miloš Ivanović[7], and Nenad Filipović[5,8]

[1]Institute for Information Technologies, University of Kragujevac, Jovana Cvijića bb, 34000 Kragujevac, Serbia
[2]BIA Separations CRO Laboratory, Teslova ulica 30, 1000 Ljubljana, Slovenia
[3]Faculty of Medical Sciences, University of Kragujevac, Svetozara Markovića 69, 34000 Kragujevac, Serbia
[4]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
[5]BioIRC - Bioengineering Research and Development Center, University of Kragujevac, Prvoslava Stojanovića 6, 34000 Kragujevac, Serbia
[6]Labena d.o.o. Serbia, Bulevar Zorana Đinđića 123G, 11070 Belgrade, Serbia
[7]Faculty of Science, University of Kragujevac, Radoja Domanovića 12, 34000 Kragujevac, Serbia
[8]Faculty of Engineering, University of Kragujevac, Sestre Janjić 6, 34000 Kragujevac, Serbia

nevena_milivojevic@live.com

Whole 3' transcriptome profiling at the single cell level opens up new abilities for researchers to answer complex questions. Thousands of individual cells per sample are Barcoded separately to index the transcriptome of each cell individually. It is done by partitioning thousands of cells into nanoliter-scale Gel Beads-in-emulsion (GEMs), where cells are delivered at a limiting dilution, such that the majority (~90-99%) of generated GEMs contain no cell. The 16 bp 10x Barcode and 12 bp UMI are encoded in Read 1, while the poly(dT) primers are used in this protocol for generating Single Cell 3' Gene Expression libraries. After GEM generation, copartitioned cells are lysed and reverse transcription (RT) was performed after which all cDNA from single cell share a common Barcode. Full-length cDNA was amplified via PCR to generate sufficient mass for library construction. This is followed by enzymatic fragmentation and size selection to optimize the cDNA amplicon size. Library construction was finished via End Repair, A-tailing, Adaptor Ligation, and PCR. P5, P7, i7 and i5 sample index, and TruSeq Read 2 (read 2 primer sequence) were added. TruSeq Read 1 and TruSeq Read 2 are standard Illumina sequencing primer sites used in paired-end sequencing. The library prepared in this way, containing the P5 and P7 primers, is ready for Illumina amplification.

Keywords: single-cell analysis, mRNA, bioinformatics, transcriptome, sequencing

Acknowledgement: This research is funded by Labena Slovenia 10xGenomics Grant Challenge (Project title: Deciphering the effects of nanosized polystyrene particles using lab-on-chip technology and transcriptome profile). This research was supported by Labena Serbia, the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract number 451-03-47/2023-01/200378 (Institute for Information Technologies Kragujevac, University of Kragujevac), 451-03-47/2023-01/200111 (Faculty of Medical Sciences, University of Kragujevac), as well as Junior projects of Faculty of Medical Sciences, University of Kragujevac JP 25/19, JP 05/20, JP 06/20 and JP 24/20. This work is supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 952603 (SGABU). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

Oral presentation

Modulating Horizontal Gene Transfer through Bistability in the Dynamics of Bacterial Restriction-Modification Systems

Marko Djordjevic*[1], Lidija Zivkovic[2], and Magdalena Djordjevic[2]

[1]Faculty of Biology, University of Belgrade, Serbia
[2]Institute of Physics Belgrade, National Institute of the Republic of Serbia, Serbia
dmarko@bio.bg.ac.rs

Restriction-modification (R-M) systems consist of genes encoding restriction enzyme and methyltransferase, often co-expressed with a specialized regulator (C protein). These systems tightly regulate their function through complex cooperative positive and negative feedback loops. R-M systems defend bacterial cells against invasion by foreign DNA, such as plasmids and bacteriophages, consequently modulating horizontal gene transfer, including transmitting pathogenic genes like antibiotic resistance determinants or virulence factors. Recent experiments have directly confirmed that the R-to-M ratio significantly impacts bacteriophage infection efficiency, rendering a subset of cells more susceptible to horizontal gene transfer.

To understand the regulatory mechanisms of R-M systems, we develop a mathematical model tightly constrained by biophysical measurements of system interaction parameters. Despite the technical complexity arising from C protein forming dimer and tetramer complexes, we analytically derive a system stability diagram that can be easily modified for various R-M system architectures. A single free parameter determines the bistability of the system, which we infer from experimental measurements across three different architectures. Surprisingly, while one class exhibits monostability, the other two demonstrate bistability.

Our model successfully explains the experimental data and reveals that modulation of the barrier to horizontal gene transfer can occur through distinct mechanisms. Bistability leads to long-lasting states susceptible to acquiring pathogenic genes, whereas stochastic fluctuations only transiently lower the transfer barrier. The precise implications of these differences for bacterial pathogenicity and evolution require further investigation. However, we propose that R-M systems capable of bistable gene expression may give rise to genetically distinct bacterial populations with potentially diverse phenotypes concerning pathogenicity and antibiotic resistance.

Keywords: restriction-modification systems, nonlinear dynamics modeling, biophysical modeling, bistability, antibiotic resistance

Acknowledgments: This work is supported by The Science Fund of the Republic of Serbia (Grant no. 7750294, q-bioBDS).

Cell-type-specific mechanistic drivers of progressive multiple sclerosis lesions

Elkjaer ML*[1,2,3], Hartebrodt A[4,8], Oubounyt M[5], Weber A[1,2,3], Vitved L[3], Reynolds R[6], Thomassen M[2,7], Rottger R[8], Baumbach J[5,8], and Illes Z[1,2,3]

[1]Department of Neurology, Odense University Hospital, Odense, Denmark
[2] BRIDGE, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[3] Department of Molecular Medicine, University of Southern Denmark, Odense, Denmark
[4] Biomedical Network Science Lab, Department Artificial Intelligence in Biomedical Engineering, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[5] Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany
[6] Department of Brain Sciences, Imperial College, London, United Kingdom
[7] Clinical Genome Center, Research Unit of Human Genetics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark
[8] Department of Mathematics and Computer Science, University of Southern Denmark, 5000 Odense M, Denmark
maria.louise.elkjaer@rsyd.dk

Understanding the drivers of compartmentalized and sustained inflammation in the brain of progressive multiple sclerosis (PMS) remains elusive. To investigate the interplay between inter- and intra-cellular molecular mechanisms in white matter (WM) lesions, we integrated single-cell transcriptome and chromatin accessibility data from PMS lesions with spatial transcriptomics of chronic active lesion borders. We identified a PMS-specific oligodendrocyte genetic program governed by the Krüppel-like factor and specificity protein (KLF/SP) gene family, implicated in myelination and stress-induced iron uptake. Additionally, we found high expression of transferrin gene (TF) and its receptor megalin (LRP2) across lesion types, suggesting autocrine communication of iron uptake potential related to iron rim lesion in smoldering MS. Additionally, inflammatory phenotype of oligodendrocytes expressing osteopontin gene and complement were observed at chronic active lesion edges. Inside the chronic active lesion, the axonal damage biomarker, neurofilament light (NFL) gene expression was upregulated, and an astrocytic-neuronal axis through fibroblast growth factor (FGF) signaling (FGFR3-FGF13) was present. Additionally, a metabolic astrocyte phenotype at the lesion border potentially segregates inflammation areas. We also identified two distinct B cell co-expression networks with different locations and gene expressions, preferring different lesion types. Overall, single-cell multi-omics enabled the identification of specific cell types with unique molecular profiles, cell-cell communications, and spatial context, contributing to lesion fate.

Keywords: white matter lesions, single-cell multi-omics, progressive multiple sclerosis, spatial transcriptomics, iron metabolism, FGF signaling, chronic active lesion

Oral presentation

AI-powered framework to predict the toxicity of microplastics

Junli Xu[1,2,3]

[1]School of Biosystems and Food Engineering, University College of Dublin, Belfield, Dublin 4, Ireland
[2]Institute of Food and Health, University College Dublin, Belfield, Dublin 4, Ireland
[3]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland
junli.xu@ucd.ie

Numerous articles have been published investigating the health effects of exposure to micro- and nanoplastics (MNPs). However, these studies have yielded inconclusive findings due to the lack of comparability between them and the complex and diverse nature of the existing toxicity data on MNPs. This study presents a predictive modeling framework for assessing the cytotoxicity of MNPs using machine learning techniques based on classification. Through a thorough literature search, a dataset comprising 1824 sample points was compiled, incorporating nine features that describe the physicochemical properties of MNPs, cell-related attributes, and experimental factors. The decision tree ensemble classifier constructed using all the features (referred to as DTE1) exhibited a high predictive accuracy of 0.95, along with a recall and precision of 0.86 each. To identify the key factors influencing the toxic properties of MNPs, feature selection was performed. A simplified classifier utilizing six influential features demonstrated a comparable performance to DTE1. These findings can guide future studies by improving experimental design and reporting practices, ultimately enhancing our understanding of the urgent health concerns related to MNPs. As more representative research data is incorporated, the developed model holds the potential for broad applicability in various settings concerning MNP cytotoxicity.

Keywords: microplastic, nanoplastic, cytotoxicity, health effect, machine learning

Acknowledgement: Funding for this research was provided by the Science Foundation Ireland (SFI)-Irish Research Council Pathway Programme Proposal ID 21/PATH-S/9290.

Newest Advances on the FeatureCloud Platform for Federated Learning in Biomedicine

Niklas Probul*[1], Mohammad Bakhtiari[1], Mohammad Kazemi Majdabadi[1], Balázs Orbán[3], Sándor Fejér[3], Supratim Das[1], Julian Klemm[1], Christina C Saak[1], Nina K Wenke[1], and Jan Baumbach[1,2]

[1] Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany
[2] Institute of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[3] Gnome Design SRL, Sfântu Gheorghe, Romania
niklas.probul@uni-hamburg.de

AI in biomedicine has been a central research topic in recent years. Although there are many different techniques and strategies, the majority rely on data that is of both high quality and quantity. Despite the steady growth in the amount of data generated for patients, it is frequently difficult to make that data useful for research because of strong restrictions through privacy regulations such as the GDPR. Through federated learning (FL), we are able to use distributed data for machine learning while keeping patient data inside the respective hospital. Instead of sharing the patient data, like in traditional machine learning, each participant trains an individual machine learning model and shares the model parameters and weights. Existing FL frameworks, however, frequently have restrictions on certain algorithms or application domains, and they frequently call for programming knowledge.
With FeatureCloud, we addressed these limitations and provided a user-friendly solution for both developers and end-users. FeatureCloud greatly simplifies the complexity of developing federated applications and executing FL analyses in multi-institutional settings. Additionally, it provides an app store that makes it easy for the community to publish and reuse federated algorithms. Apps can be chained together to form pipelines and executed without programming knowledge, making them ideal for flexible clinical applications. Apps on FeatureCloud can receive certification from both internal and external reviewers to guarantee safety. FeatureCloud effectively separates local components from sensitive data systems by utilizing containerization technology, making it robust to execute in any system environment and guaranteeing data security. To further ensure the privacy of data, FeatureCloud incorporates privacy-enhancing technologies and complies with strict data privacy regulations, such as GDPR.

Keywords: federated learning, biomedicine, privacy-preserving machine learning, patient privacy

Acknowledgement: The FeatureCloud project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 826078. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. This work was developed as part of the FeMAI project funded by the German Federal Ministry of Education and Research (BMBF) under grant number 01IS21079. This work was further funded by the German Federal Ministry of Education and Research (BMBF) under grant number 16DTM100A.

Oral presentation

Deciphering key regulatory networks and drug repurposing candidates through scRNAseq data analysis using SCANet

Mhaned Oubounyt*[1], Jan Baumbach[1], and Maria L. Elkjaer[1]

[1] Institute for Computational Systems Biology,
University of Hamburg, Hamburg, Germany
mhaned.oubounyt@uni-hamburg.de

Differences in co-expression networks between two or multiple cell (sub)types across conditions is a pressing problem in single-cell RNA sequencing (scRNA-seq). A key challenge is to define those co-variations that differ between or among cell types and/or conditions and phenotypes to examine small regulatory networks that can explain mechanistic differences. To this end, we developed SCANet, an all-in-one Python package that uses state-of-the-art algorithms to facilitate the workflow of a combined single-cell GCN and GRN pipeline including inference of gene co-expression modules from scRNA-seq, followed by trait and cell type associations, hub gene detection, co-regulatory networks, and drug-gene interactions. To illustrate the power of SCANet, we examined data from two studies. First, we identify the drivers of the mechanotype of a cytokine storm associated with increased mortality in patients with acute respiratory illness. Secondly, we find 20 drugs for 8 potential pharmacological targets in cellular driver mechanisms in the intestinal stem cells of obese mice. SCANet is available as a free, open source, and user-friendly Python package that can be easily integrated in systems biology pipelines.

Keywords: small single cell networks, GRN, GCN, mechanotyping, drug repurposing

From protein-protein to isoform-isoform interactions: the toolkit to map alternative splicing to interactome

Olga Tsoy*[1], Zakaria Louadi[2], Chit Tong Lio[2], Jan Baumbach[1], Olga Kalinina[3,4], Alexander Gress[3,4], Tim Kacprowski[5,6], and Markus List[2]
[1]University of Hamburg, Notkestrasse 9, Hamburg, Germany
[2]Technical University of Munich, Maximus-von-Imhof-Forum 3, Freising, Germany
[3]Helmholtz Centre for Infection Research, Inhoffenstraße 7, Saarbrücken, Germany.
[4]Saarland University, Homburg, Germany
[5]PLRI of TU Braunschweig and MHH, Rebenring 56, Brunswick, Germany
[6]BRICS, TU Braunschweig, Rebenring 56, Brunswick, Germany
olga.tsoy@uni-hamburg.de

Alternative splicing (AS) can impact protein structure and lead to protein-protein interaction (PPI) rewiring. Available PPI networks neglect alternative splicing isoforms: as interactions might happen only between a subset of isoforms, the PPI network contains both false-positive and false-negative interactions. Since it is not feasible to validate all isoform-isoform interactions experimentally, we present a set of tools to investigate AS on a network level: DIGGER to map splicing to the PPI network, as well as NEASE and Spycone to evaluate the functional consequences of network rewiring.
DIGGER (https://exbio.wzw.tum.de/digger) integrates PPIs, domain-domain, and residue-level interactions - the structures that might be spliced in or out and result in interaction gain or loss. Users can explore possible rewiring for an isoform or exon of interest and extract relevant subnetworks. NEASE (https://github.com/louadi/NEASE) identifies pathways that are significantly affected by network rewiring. NEASE extends classic gene set enrichment analysis by considering isoform-specific interactions affecting pathways. Spycone (https://github.com/yollct/spycone) addresses the time-course changes in AS. It searches for isoforms that demonstrate similar temporal splicing patterns and reflect the splicing co-regulation. Spycone further integrates gene set, network, and splicing-aware NEASE enrichment.
Overall, we offer a splicing-focused network analysis toolkit that allows for studying the mechanistic consequences of AS.

Keywords: bioinformatics, protein-protein interactions, alternative splicing, network enrichment, time series analysis

Oral presentation

Drugst.One - A plug-and-play solution for online systems medicine and network-based drug repurposing

Andreas Maier*[1], Michael Hartung[1], The Drugst.One Initiative, and Jan Baumbach[1,2]

[1] Institute for Computational Systems Biology,
University of Hamburg, Hamburg, Germany
[2] Computational Biomedicine Lab, Department of Mathematics
and Computer Science, University of Southern Denmark, Odense, Denmark
andreas.maier-1@uni-hamburg.de

In recent decades, the development of new drugs has become increasingly expensive and inefficient, and the molecular mechanisms of pharmaceuticals often remain poorly understood. In response, numerous computational systems and network medicine tools have been developed to prioritize drug repurposing candidates. However, such tools often require local installation and configuration or lack follow-up visual network mining capabilities. To address these challenges and simplify network exploration and drug repurposing candidate prediction, we have developed Drugst.One. It is a customizable plug-and-play solution with its own data warehousing system integrating multiple interaction databases to enable interactive modeling and analysis of the associations between proteins, drugs, and diseases. With just three lines of code, it has the capacity to convert any systems medicine software into an interactive web tool for identifying drug repurposing candidates, thus providing a powerful and accessible resource for advancing drug discovery efforts. To demonstrate the utility of Drugst.One's low-code approach, we have integrated it with 20 existing computational systems medicine tools of various types, with the intent to expand the *Drugst.One Initiative* with additional collaboration partners.
Drugst.One is, to our knowledge, the first approach to unify and simplify web-based network-based visualization and drug repurposing, posing a valuable resource for the research community. Learn more about Drugst.One and the *Drugst.One Initiative* at https://drugst.one.

Keywords: Drug repurposing, Systems medicine, Interactive network enrichment, Biomedical network exploration, Network integration, Biomedical data analysis, Data visualization

Acknowledgement: REPO-TRIAL: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 777111. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains.
RePo4EU: This project is funded by the European Union under grant agreement No. 101057619. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of "CLINSPECT-M" (grant FKZ161L0214A).
JB was partially funded by his VILLUM Young Investigator Grant nr.13154.
Collaborations partners of the Drugst.One Initiative have received the following additional funding:
This work was also partly supported by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract No. 22.00115.
This work was supported by the Technical University Munich – Institute for Advanced Study, funded by the German Excellence Initiative. This work was supported in part by the Intramural Research Programs (IRPs) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 422216132.
This project has received funding from the European Research Council (ERC) Consolidator Grant 770827 and the Spanish State Research Agency AEI 10.13039/501100011033 grant number PID2019-105500GB-I00.
IJ was supported in part by funding from Natural Sciences Research Council (NSERC #203475), Canada Foundation for Innovation (CFI #225404, #30865), Ontario Research Fund (RDI #34876), IBM and Ian Lawson van Toch Fund.
SL has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965193 for DECIDER.

Oral presentation

Fatty Acid Data Analysis Unravels Skeletal Site and Age-Specific Features of Human Bone Marrow Adiposity

Drenka Trivanović*[1], Jovana Kovačević[2], Aleksandra Arsić[1], Marko Vujačić[3], Nikola Bogosavljević[3], Ivana Okić Djordjević[1], Milena Živanović[1], Slavko Mojsilović[1], Mirjana Maljković[2], and Aleksandra Jauković[1]

[1] Institute for Medical Research, National Institute of Republic of Serbia, University of Belgrade, Dr. Subotića 4, 11000 Belgrade, Serbia
[2] Faculty of Mathematics, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[3] Institute for Orthopedy Banjica, Mihaila Avramovića 28, 11000, Belgrade, Serbia
drenka.trivanovic@imi.bg.ac.rs

As adipose tissue (AT) undergoes metabolic reprogramming with age, we investigated skeletal site-specific and age-dependent lipid profile of bone marrow adipose tissue (BMAT). Acetabular and femoral BMAT, and gluteofemoral subcutaneous adipose tissue (gfSAT) were obtained from matched osteoarthritis patients. Patients were classified into two groups: younger (≤ 60 years) and aged (>60 years) adults. BMAT and gfSAT were explored by using thin layer/gas chromatography coupled with cellular and molecular assays. Data were interpreted and visualized by applying linear discriminant analysis (LDA) and hierarchical clustering of fatty acid (FA) composition. Statistics was estimated by non-parametric tests and Spearman's rank correlation.

Analyses of total lipids revealed significantly reduced triglyceride content in femoral (fBMAT) than in acetabular BMAT (aBMAT) and gfSAT. Frequencies of spontaneously released saturated palmitic (C16:0) and stearic acids (C18:0) were higher in fBMAT than in aBMAT and gfSAT ($p=0.036$ and $p=0.046$, $n=8$). Cluster heatmap and LDA showed that fBMAT differed to acetabular and gfSAT, while acetabular and gfSAT were more similar in FA profiles. FA profiles of AT depots varied with patient's age. Contribution of palmitic acid was increased in aged group in all AT depots, while stearic acid declined in aged group in BMAT compartments only. fBMAT cellularity declined with age ($r=-0.675$, $n=14$, $p=0.037$). Additionally, the presence of $CD45^{-}CD31^{-}CD34^{+}CD24^{+}$ adipogenic progenitor (stem) cells was increased in fBMAT (0.46±0.03%) when compared to aBMAT (0.21±0.01%) depot. Femoral mesenchymal stem cells displayed pronounced adipogenesis comparing to their acetabular counterparts.

Our findings suggest that specific lipid profile of fBMAT imposes adipogenic commitment of stem cells within this skeletal site.

Keywords: bone marrow adipose tissue, fatty acids, stem cells, adipogenesis, aging

Acknowledgement: Work is supported by the Ministry of Education, Science and Technological Development of Republic of Serbia [contract number 451–03-68/2022–14/200015 with Institute for Medical Research University of Belgrade, National Institute of Republic of Serbia]

Exploration of Pharmacogenomic Biomarkers in Chronic Immune Diseases Using Single-Cell RNA Sequencing

Mario Gorenjak[1], Larisa Goričan[1], Boris Gole[1], Uršula Prosenc*[2], Erik Melén[3], Michael Kabesch[4], Anke H Maitland-van der Zee[5,6], Susanne Reinartz[7], Susanne J H Vijverberg[5,6], Uroš Potočnik[1,8,9] and the PERMEABLE consortium

[1]University of Maribor, Faculty of Medicine, Centre for Human Molecular Genetics and Pharmacogenomics, Slovenia
[2]BIA Separations CRO – Labena d.o.o., Teslova 30, 1000 Ljubljana, Slovenia
[3]Karolinska Institutet, Department of Clinical Sciences and Education Södersjukhuset, Stockholm, Sweden
[4]University Children's Hospital Regensburg (KUNO), Department of Pediatric Pneumology and Allergy, Regensburg, Germany
[5]University of Amsterdam, Amsterdam University Medical Centers, Department of Respiratory Medicine, Amsterdam, The Netherlands
[6]University of Amsterdam, Amsterdam University Medical Centers, Department of Pediatric Pulmonology, Amsterdam, The Netherlands
[7]Tergooi Medical Center, Department of Otorhinolaryngology, Hilversum, The Netherlands
[8]University of Maribor, Faculty for Chemistry and Chemical Engineering, Laboratory for Biochemistry, Molecular Biology and Genomics, Maribor, Slovenia
[9]Maribor University Medical Centre, Department for Science and Research, Maribor, Slovenia
ursula.prosenc@biaseparationscro.com

Biological therapies have revolutionized management of the severe cases of Chronic Immune Diseases refractory to the standard therapies. However, many patients do not respond to the selected biological therapy, loose response over time, or develop adverse effects. A personalized approach to treatment of these patients, based on reliable biomarkers is thus clearly needed.

Non-invasive approaches, such as use of the peripheral blood immune cells, are favored for novel biomarker discovery. However, the attention has shifted away from the bulk immune cells and towards specific immune cell sub-populations. Thus, the single-cell RNA sequencing (scRNA-seq) can prove highly valuable. By simultaneously capturing and profiling all the cells in a sample, scRNA-seq allows the analysis of cellular heterogeneity and gene expression in all immune cell sub-populations, targeted or adversely affected by the biological treatment.

In our ongoing research, scRNA-seq was utilized to analyze samples from Inflammatory Bowel Disease and Childhood Asthma patients with varied response to the biological therapy. Confounding effects of disease conditions and (biological) therapies on marker genes were eliminated using computational integration in order to identify conserved marker genes across all states. It turned out, that a reliable identification of the different immune cell sub-populations in this setting is quite challenging due to subjective cell-landscape clustering resolution. Several resolutions and automated annotation approaches were subsequently tested and validated.

A reference-based approach (Seurat-Azimuth) combined with manual cluster validation proved superior. Alas, manual cluster validation is time consuming. Annotation validation is important, especially to provide additional insights into unidentified clusters, which are essential for the identification of predictive biomarkers for personalized therapies in the vast heterogeneity of immune cell landscapes residing behind pathophysiology of chronic immune diseases.

Keywords: precision medicine, chronic immune diseases, biological therapy, Single-Cell RNA Sequencing, identification of cell sub-populations

Acknowledgments: This research was funded by the Slovenian Research Agency-Research core funding P3-0427; Research grant J3-9258; the Labena company with Grant challenge program and the PERMEABLE consortium.
The PERMEABLE consortium is supported by the ZonMW (project number: 456008004), the Swedish Research Council (proj.nr. 2018-05619), the Ministry of Education, Science and Sport of the Republic of Slovenia (proj.nr. C3330-19-252012), and the German Ministry of Education and Research (BMBF) (proj.nr. FKZ01KU1909A), under the frame of the ERA PerMed JTC 2018 Call.

Dehydrins in the service of protecting the DNA helix from the aspect of molecular dynamics (MD)

Milan Senćanski*[1], Ivana Prodić[1], Ana Pantelić[1], and Marija Vidović[1]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
milan.sencanski@imgge.bg.ac.rs

Drought stress is one of the greatest threats to global food security, posing a major challenge to agriculture. Understanding the molecular mechanisms underlying desiccation tolerance in resurrection plants like *Ramonda serbica* Panc., can provide valuable insights for improving crop resilience. Dehydrins are intrinsically disordered proteins known to accumulate in these plants in response to desiccation. Among several proposed physiological roles, it has been suggested that dehydrins can protect DNA from damage during water shortage. Here, we have characterised dehydrins from *R. serbica*, selected a representative one and evaluated its potential to interact with DNA.

Most of the *R. serbica* dehydrins were designated as hydrophilins (glycine content >6%; GRAVY index <1). They exhibit a high disorder propensity, making them quite dynamic in solution. Furthermore, they were predicted to localize in the nucleus. To examine the potential interactions with DNA *in silico*, we have selected a representative, highly hydrophilic dehydrin (Gravy index: −1.29) containing a high percentage of glycine (22.6%) and charged amino acids (lysine, glutamate and aspartate). Its 3D structures were determined using the Phyre 2 intensive modelling and AlphaFold.

The dehydrin-DNA complex was manually adjusted, following molecular dynamic simulation (MDS) in both cases of hydration and desiccation. To simulate complete hydration, the DNA-dehydrin complex was solvated in a water box, with final dimensions of 100×69×82 Å, neutralised with 0.15 M NaCl. The system underwent a 10,000-step energy minimization, consecutive 1250 ps equilibration NVE (constant number of atoms, volume and energy) heating from 10 K to 298 K and 100 ns NPT (constant number of atoms, pressure and temperature) MD production at 1 bar, and 1 fs integration step. In all simulations, periodic boundary conditions (PBC) were implemented and the CHARMM36 force field was used. The obtained results revealed that selected dehydrin can interact with both minor and major DNA grooves. The phosphate groups from the DNA molecule form salt bridges with the positively charged lysines from polylysine, K-segment, contributing to the complex stability. Overall, we have provided evidence for possible dehydrin-DNA interactions. However, the exact nature and significance of these interactions is still an area of active research *in vitro*.

Keywords: intrinsically disordered proteins, dehydrins, DNA, drought, molecular dynamics, *Ramonda serbica*.

Acknowledgement: This research was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, grant number 451-03-47/2023-01/200042.

Oral presentation

Machine learning approach in inferring main population-level COVID-19 risk factors

Sofija Marković*[1], Anđela Rodić[1], Ognjen Milićević[2], Igor Salom[3], Magdalena Đorđević[3], and Marko Đorđević[1]

[1]Faculty of Biology, University of Belgrade, Studentski trg 6, 11000 Belgrade, Serbia
[2]School of Medicine, University of Belgrade, dr Subotića starijeg 8, 11000 Belgrade, Serbia
[3]Institute of Physics Belgrade, National Institute of the Republic of Serbia, University of Belgrade, Pregrevica 118, 11000 Belgrade, Serbia
sofija.markovic@bio.bg.ac.rs

Machine-learning methods have become indispensable in scientific research as the amount of available data has grown exponentially in recent years. It is, thus, necessary to employ various unsupervised and supervised machine learning methods to uncover the main determinants of COVID-19 transmissibility and severity in the population. Upon introducing appropriate disease transmissibility and severity measures and gathering relevant socio-demographic, environmental, and health-related data for the countries with obtained said measures, we implement several machine-learning-based approaches to select the most prominent drivers of disease transmissibility and severity. These approaches include regularization-based linear regression models and more advanced Random Forest and Gradient Boost methods, which are not limited to the linear relationships between the features and the response. Principal component analysis was used for preselection to avoid overfitting, where numerous features were considered for a relatively small number of observations (i.e., countries/states). As a result, a broad range of potential COVID-19 risk factors was reduced to several prominent features, selected robustly by different methods - we further untangle how they, directly or indirectly, contribute to the transmissibility and severity of the disease. Our results underscore the evolving nature of COVID-19, from the severity experienced during the first wave to the emergence of new, highly transmissible variants like Omicron. These insights can guide public health interventions, vaccine strategies, and policies aimed at reducing the burden of COVID-19 and effectively managing future waves and emerging variants.

Keywords: COVID-19, machine learning, ecological regression analysis, epidemiological modeling, outburst risk factors

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

Using whole exome sequencing to explore genetic basis of unicuspid aortic valve disease

Martina Mia Mitić[1], Dušan Ušjak*[1], Maja Milošević[2], Marija Cumbo[1], Sofija Dunjić Manevski[1], Branko Tomić[1], Ivana Petrović[2], Petar Otašević[2], Slobodan Micović[2], Milovan Bojić[2], and Valentina Đorđević[1]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia
[2]Institute for Cardiovascular Diseases Dedinje, Heroja Milana Tepića 1, Belgrade, Serbia
valentina@imgge.bg.ac.rs

Normal aortic valve consists of three cusps that develop in the embryonic stage. Unicuspid aortic valve (UAV) is a rare congenital anomaly resulting in only one cusp with estimated prevalence of 0.02% in general population. Aim of this study was to identify genetic variants possibly associated with development of UAV. The study included 17 subjects, namely 5 UAV patients and their healthy family members without UAV disorder. Total DNA was isolated from venous blood samples and whole exomes sequencing (WES) was performed using BGI's WES protocol. Adapter-trimmed and quality-filtered reads (fastp) were mapped to hg38 reference genome using BWA/SAMtools. VCF files were generated using GATK (BaseRecalibrator, HaplotypeCaller) and annotated with InterVar and AnnoVar tools. Rare heterozygous variants present in UAV patients were found in NOTCH1, TGFB2, MYH6, EGFR, FBN2, C1R, ROBO4 and TBX5, genes associated with development of aortic valves. Among these, most were missense mutations with damaging effects as predicted using *in silico* tools (SIFT and/or Polyphen). Only mutation in MYH6 p.Ala1130Ser was found in at least two different UAV patients. Also, rare homozygous missense mutation p.Gly577Ser with high damaging potential was found in ADAMTS5 gene. Besides, highly damaging heterozygous missense mutations were detected in gene interacting functional partners (STRING) of genes associated with development of aortic valves: DVL1, THBS1, NOTCH4, ADAMTS3, FBN1, NOTCH2, ADAM17, LRP5, WWTR1, C1S, ANKRD6 and TNNI1, as well as homozygous in ACAN and KNG1. Taken together, malfunctions in ADAMTS5, ACTA2, MYH6, FBN2, AXIN1, CELSR1 or TBX5 networks were found to be common in at least two UAV patients, suggesting existence of genetic basis in UAV disorder, possibly as a result of combined effects of multiple variants.

Keywords: unicuspid aortic valve, congenital heart disease, whole exome sequencing, genetic variants, valves

Oral presentation

Online *in silico* validation of disease and gene sets, clusterings or subnetworks with DIGEST

Klaudia Adamowicz*[1], Andreas Maier[1],
Jan Baumbach[1,2], and David B. Blumenthal[3]

[1]Institute for Computational Systems Biology, University of Hamburg, Hamburg, Germany
[2]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
[3]Department Artificial Intelligence in Biomedical Engineering (AIBE), Friedrich-Alexander-Universitaet Erlangen-Nuernberg (FAU), Erlangen, Germany
klaudia.adamowicz@uni-hamburg.de

Given the constraints faced in the development of new drugs, the importance of drug repurposing has reached unprecedented levels. A key aspect of effective drug repurposing lies in the discovery of disease mechanisms and the identification of clusters of diseases with shared mechanistic characteristics. While various methods exist for computing candidate disease mechanisms and clusters, the absence of ground truth presents challenges in validating these predictions through *in silico* means. This obstacle significantly impedes the widespread adoption of *in silico* prediction tools, as experimentalists often hesitate to conduct wet-lab validations without clearly quantified initial plausibility.

To address this issue, we introduce DIGEST (*in silico* validation of disease and gene sets, clusterings or subnetworks). DIGEST is a Python-based validation tool that offers multiple avenues for utilization. It is accessible as a web interface through https://digest-validation.net, as a stand-alone package, or via a REST API. DIGEST streamlines the process of *in silico* validation by providing fully automated pipelines. These pipelines encompass critical components such as disease and gene ID mapping, enrichment analysis, comparisons of shared genes and variants, and background distribution estimation. Additionally, DIGEST incorporates functionality to automatically update the external databases utilized by the pipelines. By employing DIGEST, users gain the ability to assess the statistical significance of candidate mechanisms in terms of functional and genetic coherence. The tool enables the computation of empirical P-values with ease, requiring only a few simple clicks. With its comprehensive and user-friendly features, DIGEST greatly facilitates the evaluation of candidate mechanisms, empowering researchers to quantify the plausibility of predicted mechanisms in a robust and efficient manner.

Keywords: Systems medicine, in silico validation, Functional and genetic coherence

Acknowledgement: This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements No. 777111 (A.M., J.B.). This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant 01ZX1908A and grant 01ZX1910D) (J.B.). J.B. was partially funded by his VILLUM Young Investigator Grant No. 13154.

Alternative splicing impacts microRNA regulation within coding regions

Lena Maria Hackl*[1], Amit Fenn[1,2], Zakaria Louadi[1,2], Jan Baumbach[1,3], Tim Kacprowski[4,5], Markus List[2], and Olga Tsoy [1]

[1] Institute for Computational Systems Biology, University of Hamburg, Notkestrasse 9, 22607 Hamburg, Germany
[2] Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany
[3] Computational BioMedicine Lab, University of Southern Denmark, Campusvej 50, 5230 Odense, Denmark
[4] Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Rebenring 56, 38106 Braunschweig, Germany
[5] Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Rebenring 56, 38106 Braunschweig, Germany
olga.tsoy@uni-hamburg.de

MicroRNAs (miRNAs) are small non-coding RNA molecules that regulate post-transcriptional gene expression by binding to specific target sites. Approximately 95% of human multi-exon genes can be spliced alternatively, which enables the production of functionally diverse transcripts and proteins from a single gene. In complex diseases, such as cancer, gene but also miRNA dysregulation plays a significant role. According to most studies miRNAs preferably bind to 3'-untranslated regions of mRNA. However, through alternative splicing, transcripts might lose exons harboring miRNA target sites and, hence, become unresponsive to miRNA regulation.

To check this hypothesis, we studied the role of miRNA target sites in both coding and noncoding regions using six cancer data sets from The Cancer Genome Atlas (TCGA). First, we predicted miRNA target sites on mRNAs from their sequence using TarPmiR. For our analysis, we focused on miRNAs whose expression was negatively correlated with gene expression (as evidence for active regulation) as well as genes that were at least moderately expressed and showed evidence of alternative splicing. We chose different subsets of transcripts to differentiate the effects of target sites in different gene regions. To check whether alternative splicing interferes with miRNA regulation, we trained linear regression models to predict miRNA expression from transcript expression. Using nested models, we compared the predictive power of transcripts with miRNA target sites to that of transcripts without target sites in the investigated gene region. For all six cancer data sets and all subsets, models containing transcripts with target sites predicted miRNA abundance significantly better.

We conclude that alternative splicing does interfere with miRNA regulation by skipping exons with miRNA target sites within the coding region.

Keywords: alternative splicing, miRNA, machine learning, nested models, cancer

Oral presentation

Using AI to design antibodies

Goran Rakočević[1,2]

[1]Absci, Vancouver, Washington, USA
[2]School of computing, Union University, Belgrade
grakocevic@absci.com

Advancements in antibody engineering are crucial for developing effective and safe therapeutic candidates. Traditional approaches often involve limited screening of sequence space, resulting in drug candidates with suboptimal binding affinity, developability, or immunogenicity. However, recent breakthroughs in deep learning and generative artificial intelligence (AI) offer promising solutions to overcome these challenges.

In our work, we utilized deep contextual language models trained on high-throughput affinity data to quantitatively predict binding of unseen antibody sequence variants. Our approach spans a wide range of binding affinities, demonstrating the potential to optimize antibody engineering. Additionally, we introduced a "naturalness" metric that measures similarity to natural immunoglobulins. We found that naturalness is associated with measures of drug developability and immunogenicity, allowing us to optimize it alongside binding affinity using a genetic algorithm.

Additionally, we explored generative AI-based antibody design, and achieved successful design of all complementarity-determining regions (CDRs) in the heavy chain of the antibodies. Our designed antibodies exhibit high binding rates, surpassing randomly sampled antibodies from the Observed Antibody Space. Moreover, these AI-designed binders display high diversity, low sequence identity to known antibodies, and favorable naturalness scores, indicating desirable developability profiles and reduced immunogenicity.

Collectively, our findings demonstrate the immense potential of deep learning and generative AI in revolutionizing antibody optimization and design. By leveraging large-scale data, predictive models, and high-throughput experimentation, we can accelerate and improve our antibody engineering capabilities. The integration of deep contextual language models and the incorporation of naturalness into the design process provide intelligent screening approaches. Similarly, the application of generative AI enables us to efficiently and precisely design antibodies from scratch, outperforming traditional methods in terms of speed and quality.

Keywords: artificial intelligence, antibody design

Acknowledgement: I would like to thank all of my coauthors and collaborators: Amir Shanehsazzadeh, Sharrol Bachas, Matt McPartlon, George Kasun, John M. Sutton, Andrea K. Steiger, Richard Shuai, Christa Kohnert, Jahir M. Gutierrez, Chelsea Chung, Breanna K. Luton, Nicolas Diaz, Simon Levine, Julian Alverio, Bailey Knight, Macey Radach, Alex Morehead, Katherine Bateman, David A. Spencer, Zachary McDargh, Jovan Cejovic, Gaelin Kopec-Belliveau, Robel Haile, Edriss Yassine, Cailen McCloskey, Monica Natividad, Dalton Chapman, Joshua Bennett, Jubair Hossain, Abigail B. Ventura, Gustavo M. Canales, Muttappa Gowda, Kerianne A. Jackson, Jennifer T. Stanton, Marcin Ura, Luka Stojanovic, Engin Yapici, Katherine Moran, Rodante Caguiat, Amber Brown, Shaheed Abdulhaqq, Zheyuan Guo, Lillian R. Klug, Miles Gander, Joshua Meier, Anand V. Sastry, Andrew Stachyra, Borka Medjo, Vincent Blay, Alexander Brown, Nebojsa Tijanic, Rebecca Viazzo, Rebecca Consbruck, Hayley Carter, Jacob Shaul, Randal S. Olson, Sean McClain, Matthew Weinstock, Gregory Hannum, Ariel Schwartz, Roberto Spreafico

Semantic unification and search of bioinformatics databases

A. Veljković*[1], and N. Mitić[1]

[1]Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade, Serbia
aleksandar.veljkovic@matf.bg.ac.rs

Analyzing biological data from various sources offers a comprehensive perspective of a domain, facilitating the identification of patterns that would otherwise be challenging or impossible to observe when focusing solely on individual biological entities. The process of linking data from different databases can present challenges due to inconsistencies in properties and identifiers assigned to the same entity across databases. Although certain databases include a range of identifiers from multiple sources, the search capabilities are restricted to exact property matching, preventing the execution of complex queries involving multiple metadata attributes.

We designed a novel data framework that aims to address these challenges by facilitating the linkage and retrieval of information from diverse interconnected biological data sources. To evaluate the effectiveness of the model, we conducted tests and created a knowledge graph using metadata extracted from five separate public datasets: DisProt, HGNC, Tantigen 2.0, IEDB, and DisGeNET. The resulting graph establishes connections between more than 17 million nodes, comprising 2.5 million distinct biological entity objects, and encompasses over 4 million relationships.

Additionally, we designed and implemented a general-purpose procedure for extracting new relationships based on semantic similarity from data transformed into the BioGraph data model.

Keywords: Bioinformatics database, semantic search, unification, BioGraph

Oral presentation

Beyond the Global Health Security Index: A Machine Learning Approach to Analyzing the Official COVID-19 Deaths and Excess Deaths Data

Andjela Rodic*[1], Sofija Markovic[1], Igor Salom[2], and Marko Djordjevic[1]

[1]Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[2]Institute of Physics, National Institute of the Republic of Serbia, Pregrevica 118, University of Belgrade, 11000 Belgrade, Serbia
andjela.rodic@bio.bg.ac.rs

The Global Health Security Index (GHSI) is designed to assess the preparedness of countries to deal with infectious disease outbreaks. However, the COVID-19 pandemic has revealed a paradoxical relationship between the GHSI and the COVID-19 mortality, with higher GHSI scores being associated with higher death rates. We aimed to explain this puzzle. To rely on an accurate and robust measure of COVID-19 severity across countries, we used our model-derived measure instead of the standard Case Fatality Rate. We employed a range of statistical learning techniques, including non-parametric machine learning methods, to identify the factors that influence COVID-19 severity in 85 countries. Also, we searched for the predictors of the largely unexplored excess mortality counts. Our results suggest that the association of higher preparedness, measured by the GHSI, with higher COVID-19 mortality may be an artifact of oversimplified statistical analyses used in published studies. In addition, it could be a consequence of misclassified COVID-19 deaths, combined with the higher median age of the population and earlier epidemics onset in countries with high GHSI scores.

Keywords: bioinformatics, modeling epidemics, machine learning, COVID-19 severity, excess deaths

Integration of differential transcriptomic and proteomic data in hydrated and desiccated leaves of *Ramonda serbica* Panc.

Marija Vidović*[1], Ilaria Battisti[2,3], Ana Pantelić[1], Dejana Milic[1], Giorgio Arrigoni[2,3,] Antonio Masi[4], and Sonja Veljović Jovanović[5]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
[2]Department of Biomedical Sciences, University of Padova, Via Ugo Bassi 58/B, 35131 Padova, Italy
[3]Proteomics Center, University of Padova and Azienda Ospedaliera di Padova, Via G. Orus 2/B, 35129 Padova, Italy
[4]DAFNAE, University of Padova, Viale dell'Università 16, 35020 Legnaro, Italy
[5]Institute for Multidisciplinary Research, University of Belgrade, Kneza Viseslava 1, 11000 Belgrade, Serbia
mvidovic@imgge.bg.ac.rs

The resurrection plant *Ramonda serbica* Panc. survives long desiccation periods and fully recovers metabolic functions within one day upon watering. We aimed to identify key candidates and pathways involved in desiccation tolerance in *R. serbica* by employing a systems biology approach, combining transcriptomics and proteomics.
A total of 68,694 differentially expressed genes (DEGs; p-value<0.005 and abs(log2FC)≥2) were obtained in *R. serbica* leaves upon desiccation. Among them, 23,935 and 26,169 genes were upregulated and downregulated in desiccated leaves (DL) and hydrated leaves (HL), respectively. By differential TMT-based proteomic analysis 1192 different protein groups were identified after filtering with at least two unique peptides per protein. In total, 229 protein groups were more abundant in HL and 179 in DL (p-value<0.05 and abs(FC)≥1.3). The majority of the DAPs and DEGs involved in photosynthesis, transport, secondary metabolism, and signaling, were less abundant in DL. On the other hand, proteins and transcripts associated with fermentation, N-metabolism, heme, protein synthesis, folding and assembly, C1-metabolism, and late embryogenesis abundant proteins, were more accumulated in DL.
A poor correlation between proteomic and transcriptomic results was detected for mitochondrial electron transport and ATP production, gluconeogenesis, glycolysis, tricarboxylic acid cycle, and enzymatic H_2O_2 scavengers due to different mRNA half-life, protein turnover, dynamic posttranscriptional and posttranslational modifications. Finally, desiccation tolerance in *R. serbica* is a species-specific process orchestrated by several metabolic pathways that are temporally and compartmentally regulated at several levels.

Keywords: differentially abundant proteins, differentially expressed genes, drought, resurrection plants

Acknowledgements: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-47/2023-01/200042) and by the Science Fund of the Republic of Serbia-RS (PROMIS project LEAPSyn-SCI, grant no. 6039663).

Poster presentation

Possible role of estrogen metabolism and aldo-keto reductase activity in chemoresistance of ovarian cancer

Nika Marolt[1], Andrew Walakira[1], Tadeja Režen[1], Damjana Rozman[1] and Tea Lanišnik Rižner[1]

[1]Institute of Biochemistry and Molecular Genetics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia

High-grade serous ovarian cancer (HGSOC) is the most aggressive and chemoresistant form of epithelial ovarian cancer (OC) and is responsible for ~80% of OC-related deaths. OC is associated with disturbed estrogen action. In postmenopausal patients, estrogens are formed locally from steroid precursors. Enzymes of the AKR1C subfamily are associated with resistance to chemotherapeutic agents and are involved in the biosynthesis and metabolism of steroid hormones, thus may contribute to the growth of hormone-dependent tumors. To date, the interplay of estrogen synthesis and aldo-keto reductase activity in HGSOC chemoresistance remains unclear.

The aim of this study was to investigate the differences in targeted transcriptomics of HGSOC cell lines with different sensitivity to carboplatin: OVSAHO, OVCAR-3, Kuramochi, OVCAR-4, Caov-3, and COV362, and to evaluate the differences in correlation patterns between targeted gene expression profiles in platinum-sensitive and -resistant patients using publicly available data (PAD) (cBioPortal).

We first determined the expression of genes involved in estrogen biosynthesis/metabolism (*STS*, *SULT1E1*, *HSD17B1*, *HSD17B2*, *HSD17B14*, *PAPSS1*, *PAPSS2*), steroid transport (*SLCO1A2*, *SLCO1B3*, *SLCO2B1*, *SLCO4A1*, *SLCO4C1*, *ABCC1*, *ABCC4*, *ABCC11*, *ABCG2*, *SLC51A*, *SLC51B*), estrogen action (*ESR1*, *ESR2*, *GPER*) and oxidative metabolism (*CYP1A1*, *CYP1A2*, *CYP1B1*, *SULT1A1*, *SULT2B1*, *SULT1E1*, *UGTB7*, *COMT*, *NOQ1*, *NOQ2*, *GSTP1*), *NFE2L2* and *AKR1C1-3* by qPCR. Next, by using PAD we conducted a correlation analysis using the Pearson correlation coefficient for gene expression data of targeted genes in OC patients. The patients were classified into two groups based on their response to platinum treatment: sensitive and resistant. The correlation matrix was computed independently for each group.

Expression analysis revealed that the estrogen receptor *ESR2,* the efflux transporter *ABCG2* and aldo-keto reductase *AKR1C1* were highly expressed in the most resistant cell lines COV362 and Caov-3. The mRNA levels of estrogen biosynthesis and oxidative metabolism genes *STS*, *HSD17B14*, *NOQ1*, and *GSTP1* increased with carboplatin resistance in the HGSOC cell lines. These results indicate the potential of *ESR2, STS*, *HSD17B14*, *NOQ1*, *GSTP1*, and *ABCG2* as predictive markers for HGSOC chemoresistance. Furthermore, analysis of PAD revealed different correlation profiles between genes in sensitive and resistant patients. In chemoresistant were found a moderately to strong positive correlations ($p<0.001$) between gene pairs including *AKR1C1–AKR1C3, AKR1C1 – NFE2L2, AKR1C1 – SULT1E1, NOQ1 – HSD17B14, COMT – SULT1A1, ABCG2 – SLC515.* In chemosensitive patients was found a strong positive correlation ($p<0.001$) between gene pair *CYP1B1 – SULT1E1.* The correlation differences between sensitive and resistant OC patients suggest possible gene regulatory networks or molecular interactions contributing to the heterogeneity of response to platinum in OC. Further studies are ongoing to elucidate the mechanism of the interplay between local estrogen metabolism and aldo-keto reductase activity in HGSOC chemoresistance.

We gratefully acknowledge dr. Maja Zagorščak for her valuable assistance in implementing the R code used in this study.

Seven miRNAs potentially included in the chilling response of maize plants in early stages of development

Manja Božić[1]*, Dragana Ignjatović-Micić[1], Nenad Delić[1], Marko Mladenović[1], Jelena Vančetović[1], Bojana Banović Đeri[2], Ana Nikolić[1]

[1] Maize Research Institute „Zemun Polje", Slobodana Bajića 1, 11085 Belgrade, Serbia
[2] Institute of Molecular Genetics and Genetic Engineering, Vojvode Stepe 444a, 11042 Belgrade, Serbia
mbozic@mrizp.rs

Micro RNAs (miRNAs) are known regulators of various processes in plants, including growth, development and stress responses. They achieve this through mRNA cleavage or translational inhibition, in a process called RNA interference. Herein, their role in chilling stress response in young maize seedlings (*Zea mays* L.) is examined, using high-throughput sequencing methods. Bringing light to all aspects of chilling stress response in maize is necessary since earlier sowing, during colder periods, is one of the most promising strategies of avoiding maize yield loss due to effects of climate change in these areas.

Sterilized seeds of two maize genotypes (tolerant - T and sensitive - S to low temperatures) were germinated in the dark for five days (optimal conditions), after which the 5-d old seedlings were exposed to chilling conditions for 6h (10° C). Samples for RNA isolation and cDNA library preparation were taken after the treatment ended, and single-end 50 bp sequencing was performed (Illumina® Novaseq 6000). The miRNAs were then filtered, mapped, identified and quantified using adequate bioinformatics tools; and the differential expression analysis was carried out using the DEGseq R package. The analysis was performed on 859 miRNAs, after previously executed TPM normalization using the MA-plot-based method with random sampling model (MARS). The threshold for significantly differential expression was set as the Bayesian adjusted p-value, or q-value < 0.01 and log2 fold change > 1.

A total of 612 were expressed differentially, but only 55 miRNAs were common for both genotypes and at the same time differentially expressed between control and treatment conditions – 40 novel and 15 known. Half of the common miRNAs showed the same expression patterns in both genotypes, while the other half did not. Among them, seven known miRNAs showed opposing expression patterns between the genotypes (zma-miR167b-3p zma-miR167e-3p, zma-miR159c-5p, zma-miR164g-3p, zma-miR166a-5p, zma-miR398a-3p, and zma-miR528a-3p). These miRNAs were shown to have a role in various abiotic stress responses, including drought, waterlogging, high salts – but not chilling. While the results point to their potential role in establishing chilling tolerance in maize seedlings, further research is necessary to confirm it and connect the miRNAs to their potential targets.

Keywords: maize, abiotic stress, chilling, high-throughput sequencing, miRNAs

Poster presentation

Two contrasting late embryogenesis abounded protein family groups of *Ramonda serbica* Panc.

Ana Pantelić[1]*, Strahinja Stevanović[1], Sonja Milić Komić[2], Nataša Kilibarda[3] and Marija Vidović[1]

[1] Laboratory for Plant Molecular Biology, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
[2] Department of Life Science, Institute for Multidisciplinary Research, University of Belgrade, Kneza Višeslava 1, 11000 Belgrade, Serbia
[3] Department of Pharmacy, Singidunum University, Danijelova 32, 11000 Belgrade, Serbia
anapantelic@imgge.bg.ac.rs

Ramonda serbica Panc. is an ancient resurrection plant, that survives a long desiccation period and fully recovers metabolic functions upon watering. The main characteristic of desiccation-tolerant plant species is their ability to accumulate protective late embryogenesis abounded protein (LEAPs). To propose their role in *R. serbica* desiccation tolerance we structurally analysed LEAPs in hydrated and desiccated leaves.

According to transcriptomics, 318 LEAPs were identified and classified into seven family groups based on protein BLAST analysis and conserved motifs (Pfam). The largest LEAPs belonged to the LEA2 and LEA4 protein family groups. We employed online tools to analyse physicochemical characteristics (Expasy, ProtParam, BioPython, GRAVY calculator), disorder propensity, and characterization protein structures (FELLS, JPred, SOPMA, PsiPred, Phyre2, Espritz-DisProt, Espritz-X, Iupred, TMHMM, +Heliquest).

The most abundant, atypical LEA2 group containing 127, mostly hydrophobic proteins, was divided into five subgroups. Members of this group were predicted to fold into globular domains, β-barrel at the C-terminus, followed by transmembrane hydrophobic-helices and disordered N-terminal regions. Results indicated the possible involvement in the protection of the chloroplastic membranes.

The LEA4 group exhibited an exceptionally high tendency to form amphipathic α-helices and simultaneously had a high disorder propensity. This group is made of 96 proteins, classified into 3 subgroups. The high content of polar and charged amino acids (lysine, glutamate, and aspartate) is characteristic of this group. Motifs corresponding to the *R. serbica* LEA4 protein family group folded into A-type α-helices that contained positive, negative, and hydrophobic surfaces. Based on previous knowledge, the possible functions of the LEA2 and LEA4 groups are discussed with significant implications on cell preservation technology and the improvement of crop drought tolerance.

Keywords: LEA proteins, secondary structure prediction, IDPs, resurrection plants

Acknowledgements: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-47/2023-01/200042) and by the Science Fund of the Republic of Serbia-RS (PROMIS project LEAPSyn-SCI, grant no. 6039663).

De novo Genome Assembly of Sweet Chestnut (*Castanea sativa* Mill.) Insights into the Molecular Basis of its Nutritional Properties

M. Aydin Akbudak[1]* and Ali Tevfik Uncu[2]

[1] Akdeniz University, Department of Agricultural Biotechnology, Antalya, Turkiye
[2] Necmettin Erbakan University, Department of Molecular Biology and Genetics, Konya, Turkiye
akbudak@akdeniz.edu.tr

The Sweet Chestnut (*Castanea sativa* Mill.) is a tree species that holds significant economic importance and naturally spreads throughout central-southern Europe and Asia Minor. Its highly nutritious nuts have a unique composition that sets them apart from other nuts, being rich in vitamins, including vitamin C, and B vitamins such as thiamine, niacin, and folate. Over the last few decades, breeding efforts have prioritized the development of sweet chestnut cultivars that are resistant to blight and produce better nuts. However, despite these efforts, molecular genetic studies of the sweet chestnut have been insufficient. To bridge this knowledge gap, we set out to create the first reference genome of the sweet chestnut using whole-genome shotgun paired-end sequencing. Our study involved genome-wide analyses to identify and functionally annotate genes in sweet chestnut, and develop and confirm SSR-SNP markers. Additionally, we have identified and characterized specific genomic loci that enhance the nutritional value of sweet chestnuts. To the best of our knowledge, this is the first study to investigate the genetic loci responsible for determining the nutritional value of chestnuts. We anticipate that our findings will significantly contribute to the development of sweet chestnut cultivars with higher levels of bioactive compounds, minerals, and digestibility, ultimately enhancing the nutritional value of chestnuts.

Keywords: Genome sequencing, sweet chestnut, genomic loci, nutritional value

Poster presentation

Numerical and Biological Modeling Approach in the Analysis of the Cancer Viability and Apoptosis

Katarina Virijević[1,5]*, Marko Živanović[1], Marina Gazdić Janković[2], Amra Ramović Hamzagić[2], Nevena Milivojević[1], Katarina Pecić[1], Dragana Šeklić[1], Milena Jovanović[3], Nikolina Kastratović[2], Ana Mirić[1], Tijana Đukić[1], Ivica Petrović[2], Vladimir Jurišić[2], Biljana Ljujić[2], Nenad Filipović[4,5]

[1] Institute for Information Technologies, University of Kragujevac, Jovana Cvijića bb, 34000 Kragujevac, Serbia
[2] Faculty of Medical Sciences, University of Kragujevac, Svetozara Markovića 69, 34000 Kragujevac, Serbia
[3] Faculty of Sciences, University of Kragujevac, Radoja Domanovića 12, 34000 Kragujevac, Serbia
[4] Faculty of Engineering, University of Kragujevac, Sestre Janjić 6, 34000 Kragujevac, Serbia
[5] Bioengineering Research and Development Center (BioIRC), Prvoslava Stojanovica 6, 34000 Kragujevac, Serbia
msc.katarina.virijevic@gmail.com

Biomedicine is a multidisciplinary branch of science that requires a clear approach to the study and analysis of various life processes necessary for a deeper understanding of human health. This research focuses on the use of numerical simulations with the aim of an improved comprehension of cancer viability and apoptosis during treatment with commercial chemotherapeutic agents. In recent times, the usage of numerical models was successfully applied to predict the behavior of tumors. This study includes a wide range of numerical results that have been obtained by examining cell viability in real-time, determining the type of cell death and the genetic factors that control these processes. The results of the *in vitro* test were used to develop a numerical model that provides a new perspective on the proposed problem. In this study, colon, and breast cancer cell lines (HCT-116 and MDA-MB-231), as well as healthy lung fibroblast cell line (MRC-5) were treated with commercial chemotherapeutic agents. The obtained results showed a decrease in viability and the occurrence of predominantly late apoptosis upon treatment, as well as a strong correlation between parameters. A mathematical model was developed and used to gain a better understanding of the investigated processes. This method can accurately simulate the behavior of cancer cells and reliably predict their growth.

Keywords: numerical modeling, cancer, cell viability, apoptosis, gene expression, cytostatics

Acknowledgment: The authors are grateful for the support of the European Union's Horizon 2020 research and innovation programme (grant agreement No 952603 (SGABU)). This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains. This research is supported by the Serbian Ministry of Education, Science, and Technological Development [451-03-9/2021-14/200378 (Institute for Information Technologies, University of Kragujevac)].

Root colonization ability of herbicide-resistant PGP bacteria evaluated by 16S rRNA metabarcoding

Cristina Bez[1], Ivana Galic[2]*, Iris Bertani[1], Nada Stankovic[2], Vittorio Venturi[1]

[1] International Centre for Genetic Engineering and Biotechnology, Padriciano 99, 34149 Trieste, Italy
[2] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
ivanagalic@imgge.bg.ac.rs

In terms of agricultural sustainability, herbicide-resistant, plant growth promoting (PGP) bacteria that can improve crop yield are valuable resource. To exhibit PGP traits, the bacteria must be able to colonize and survive in the rhizosphere.
Upon screening the herbicide-resistant bacterial collection, candidates with the highest PGP potential were grouped into three consortia to evaluate their ability to colonize roots and persist in the natural/local plant microbiome in the pot. Experiments were conducted with seeds of commercial maize hybrids under controlled conditions, with and without herbicide. Colonization ability was evaluated by examining multiple plants from each treatment at two-time points during the experiment. 16S rRNA amplicon community profiling was performed to precisely target the bacterial strains used in the three consortia and investigate how the local microbiome might be altered by the application of the consortia. Bioinformatic analysis was performed using qiime2, clustering of reads into amplicon sequence variants ASVs using the DADA2 plugin, and the taxonomic assignment was based on a customized dataset formed from the 16S rRNA gene sequences of the ten isolates used in this study or by using the Silva rRNA database. For clustering and comparison of ASVs based on sequence similarity, the program cd-hit was used, with the sequence similarity parameter set to 98% to be considered part of the same cluster. The obtained dataset was imported into R using the package qiime2R, and subsequent analyzes and graphs were generated using either the R packages phyloseq, microbiome, or reshape2. We identified seven out of ten inoculated strains in both time points tested and with comparable abundance, indicating that most of the bacterial isolates tested have the ability to colonize the root system of maize. Furthermore, the natural/local microbiome of maize plants is not disturbed by the three consortia used in this study, implying that they are good candidates for future biotechnological applications.

Keywords: metabarcoding, 16S, consortium, bacteria, PGP

Acknowledgement: FEMS Research and Training Grant ID: 1818; IMGGE work program for 2022 (Ministry of Education, Science and Technological Development of the Republic of Serbia, 451-03-68/2022-14/200042)

Poster presentation

Genetic Complexity and Synteny Analysis of Castanea Genomes: Unveiling the Significance of Chestnut Species in Ecological and Genomic Perspectives

Ali Tevfik Uncu[1] and M. Aydin Akbudak[2]*

[1] N. Erbakan University, Department of Molecular Biology and Genetics, Konya, Turkiye
[2] Akdeniz University, Department of Agricultural Biotechnology, Antalya, Turkiye
akbudak@akdeniz.edu.tr

Castanea, a prominent genus within the *Fagaceae* family, thrives across the expansive woodlands of eastern North America, Europe, and Asia, holding considerable ecological and economic significance. Among the invaluable forest resources, chestnuts play a pivotal role by providing both nourishment and wood products. Furthermore, they assume the status of keystone species due to their indispensable ecological functions in afforestation and the provision of crucial ecosystem services. The genomes of *C. mollissima* and *C. sativa*, estimated to be around 800 Mb in size, add to the remarkable genetic complexity of these species. We utilized the powerful Python module jcvi to conduct a thorough synteny analysis, focusing on the Chinese and European chestnut genomes. To detect structural variants between these genomes, we employed SyRI (Synteny and Rearrangement Identifier). Additionally, TBtools was utilized to visually illustrate the syntenic genes across different genomes. Our investigation involved a comprehensive synteny analysis of the Chinese chestnut genome and the Sariaslama cultivar of the European chestnut. Encouragingly, we observed a strong overall synteny between these genomes, indicating significant conservation. To enhance the accuracy and completeness of the genome assemblies, we employed Pacbio sequencing technology, which contributed to the high-quality results obtained for both the European and Chinese chestnut genomes.

Keywords: Genome sequencing, European chestnut, Chinese chestnut, Synteny analysis

Acknowledgement: The study was funded by Akdeniz University Scientific Research Projects Coordination Unit. Grant number: FBA- 2023-6237.

Elongation factor P (-like) protein and polyproline motifs

Marina Parr[1*], Alina Sieber[2], Prof. Dr. Dmitrij Frishman[1] and Dr. Jürgen Lassak[2]

[1] Technical University of Munich, Germany
[2] Ludwig-Maximilians-Universität München, Germany
mar.ark.parr@gmail.com

Two or more consecutive prolines induce ribosome stalling during translation. In bacteria the elongation factor P (EF-P) efficiently rescues the ribosome stalling and allows the protein biosynthesis to continue. A seven amino acids long loop between beta-strands β3/β4 is crucial for EF-P function. The residue at the tip of the loop is subjected to the post-translational modifications: lysine is lysylated or arginine is rhamnosylated. We have demonstrated that only those enzymes that are needed for specific post-translational modification of the tip are coded in the bacterial genome (EpmA, EpmB and EpmC proteins for EF-P with lysine and EarP- for those with arginine). Phylogenetic analysis has also unveiled an invariant proline in the -2 position of the tip of the loop in EF-Ps that utilize lysine modifications such as *Escherichia coli*. Bacteria with the arginine modification like *Pseudomonas putida* on the contrary have selected against it. Combining these observations with experimental evidence, we conclude that β3/β4 loop composition is important for functionalization of EF-P by chemically distinct modifications.
Some bacterial genomes also code the elongation factor P-like (EfpL) protein that shares the same domain architecture with EF-P and has an extended loop of eight amino acid residues long. The evolution, sequence and the structure of EfpL protein have been extensively characterized. Using the assay based on luminescence emission and ribosomal profiles we have shown that EfpL can also relieve the arrest of the ribosome induced by polyproline motifs.
We have also observed the negative correlation between the occurrence of the motif in the proteome of *Escherichia coli* and its stalling strength measured in luminescence assay. We hypothesize that motifs that cause strong ribosome stalling are disfavored in the protein sequences during evolution due to their impact on the dynamics of translation.

Keywords: polyproline motifs, translation, post-translational modifications, evolution, ribosome profiling, ribosome stalling

Poster presentation

Comparative study of *in silico* protein design techniques

Ivan Tanasijević[1]*, Branka Rakić[1]

[1] Institute for Artificial Intelligence R&D, Fruskogorska 1,
21000 Novi Sad, Serbia
ivan.tanasijevic@ivi.ac.rs

Protein design plays a pivotal role in various scientific and industrial applications, such as drug development and biotechnology. With the advancement of computational methods, new tools and algorithms have emerged to facilitate the generation of novel protein designs. This study presents a comparative analysis of Pepspec and RFdiffusion, two prominent methods in protein design, to evaluate their effectiveness in designing peptides with desired properties. Mainly, we aim to design peptides that bind with high affinity and specificity to a desired protein target.

Pepspec is an application native to the Rosetta software package. It relies on Monte Carlo sampling of backbone confirmations and residue mutations and a stochastic optimization based on the Rosetta score – a measure approximating the binding free-energy of the complex.

On the other hand, a recently developed tool, RFdiffusion, is a denoising diffusion probabilistic model based on an existing artificial neural network, RoseTTAFold, developed for protein structure estimation. It is trained to remove noise from protein structures on a large database of protein complexes to ultimately be able to generate novel binder designs based on the target structure.

In this study, we aim to compare the efficiency of these two design tools. As it is common in generative ML algorithms, the comparison will be made by evaluating both the design quality and design versatility. The quality will be assessed by using the well-known AlphaFold2 Machine learning tool to estimate the binding affinity of the peptide-protein complex while the versatility will be measured using standard sequence based statistical methods.

RFdiffusion and Pepspec offer distinct approaches to protein design. By assessing the strengths and limitations of each method in this study, we aim to deepen the understanding of these methods and allow leveraging these tools effectively in designing peptides with desired characteristics, contributing to advancements in the field of protein engineering and biotechnology.

Keywords: Rational protein design, AI/ML in biology and medicine, Computational bioengineering

Energy and information exchange between "donor" and "molecular bridge" structures: non adiabatic polaron model

Dalibor Chevizovich[1], Vasilije Matic[1]*, and Zeljko Przulj[1]

[1] "Vinča" Institute of Nuclear Sciences, National Institute of the Republic of Serbia, University of Belgrade P.O.Box 522, 11001 Belgrade, Serbia
cevizd@vin.bg.ac.rs

Molecular chains (such as protein chains with alpha-helical secondary structure, DNA and RNA molecules) can play the role of "bridges" for the highly efficient transfer of various types of submolecular excitations (vibron excitations or electrons) over very long distances (comparable to the length of the molecular chain itself). In the case when this process takes place in living cells, the biomolecule is placed in an environment where it is usually in thermodynamic equilibrium with the "heat bath". As a result, the structural elements of the molecular chain perform mechanical oscillations. In the general case, such mechanical oscillations disrupt the ability of the molecular bridge to transfer the excitation over a longer distance.

On the other side, by interacting with the thermal oscillations of the structure, excitations injected into the molecule may be trapped and can form a stable self-trapped (polaron-like) state. Such quasiparticles can move through the structure with minimal energy loss. In this way, the high efficiency of energy and charge transport in living cells can be explained. However, the properties of the possibly formed polaron quasiparticle must also be affected by the presence of the donor molecule.

Here, we have discussed the mechanism of excitation transfer from a molecular structure (donor molecule) to the molecular chain. The presence of the donor structure and the temperature influence on the energy of the self-trapped excitation were considered in the dependence of the basic energy parameters of the molecular bridge. The obtained results indicate the possibility of the formation of two types of self-trapped states: a quasi-free excitation, which can easily move through the molecular bridge, and a localized, practically immobile excitation, which is similar to a non-adiabatic polaron quasiparticle.

Keywords: energy transfer, information transfer, biomolecular structures, self-trapping, polaron quasiparticle

Acknowledgment: Supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia and by the Project within the Cooperation Agreement between the JINR, Dubna, Russian Federation and Ministry of Education and Science of the Republic of Serbia.

Poster presentation

Profiling Pre-Replication Complex Mutations in Cancer

Jelena Kusic Tisma[1]*, Marija Orlic Milacic[2], Quang Trinh[2], Rhea Ahluwalia[2,3], Lincoln D. Stein[2,3]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia
[2] Ontario Institute for Cancer Research, 661 University Avenue, Suite 510, Toronto, ON, Canada, M5G 0A3
[3] Department of Molecular Genetics, University of Toronto, Medical Science Building, Room 4386, 1 King's College Circle, Toronto, ON, Canada, M5S 1A8
jkusic@imgge.bg.ac.rs

The pre-replication complex (preRC) consists of 15 proteins that mark DNA replication initiation sites and regulate replication timing. Deficiency in preRC proteins results in genomic instability (re-replication) and developmental defects (Meier-Gorlin syndrome). Our aim was to assess the scope of preRC gene aberrations in cancer. Variations in preRC genes were studied using CBio Portal software and TCGA PanCancer dataset. The functional impact of detected variants was evaluated in silico by three different prediction tools: SIFT (sequence and evolutionary conservation - based), PolyPhen2 (protein sequence and structure – based) and MutPred2 (supervised learning method based on neural networks).

No mutational hotspots were observed in any of the 15 preRC genes and no mutual exclusivity between mutations in preRC genes were detected. The highest alteration incidence in preRC genes was found in endometrial carcinoma and melanoma. The majority of the variations seen in preRC genes were non-synonymous. The functional assessment has shown that 253/1215 (21%) preRC gene mutations were predicted to be pathogenic with high confidence by 2/3 computational algorithms. None of the variants reached the high confidence pathogenicity score by all 3 prediction tool. In contrast, 49% of variants were predicted to be either benign by all three tools or benign by 2/3 or 1/3 tools, with the remaining 1/3 or 2/3, respectively, classifying them as low confidence pathogenic.

These finding suggest that mutations in preRC proteins might be passenger mutations and that cancer cells can tolerate them. The future step is to see whether incidence of coding vs. noncoding preRC mutations correlates with Tumor Mutation Burden (TMB) and Genome Instability Index (GII) of cancer.

Keywords: preRC, data mining, cBioPortal

Combined experimental and theoretical study of Type-II toxin-antitoxin system response to antibiotics

Bojana Ilic[1], Marko Đorđević[2], Hong-Yu Ou[3]

[1]Institute of Physics Belgrade, National Institute of the Republic of Serbia, Serbia
[2]Faculty of Biology, University of Belgrade, Serbia
[3]Shanghai Jiao Tong University, Shanghai, China
bojanab@ipb.ac.rs

Bacterial Type-II toxin-antitoxin (TA) systems, including kacAT in *Klebsiella pneumoniae*, respond to antibiotics. We investigated kacAT's regulation relevant to antibiotic persistence, which refers to the survival of antibiotic exposure by dormant bacterial cells. Elevated toxin levels may induce dormancy. KacAT complex binds and represses the kacAT promoter cooperatively, leading to highly non-linear negative feedback. Antibiotics increase transcription of the kacA and kacT genes by inducing KacA degradation and consequently reducing the KacA:KacT ratio. Our model reproduced experimental findings, explaining increased kacAT transcription and reduced [KacA]:[KacT] ratio. Interestingly, KacAT overexpression induces antibiotic stress tolerance, while deleting kacAT has no effect, which our model can also explain. KacAT, therefore, cannot induce spontaneous (in the absence of antibiotics) persister formation. Earlier theoretical models, which predicted spontaneous persistence in Type-II TA systems, assumed the cooperative action of multiple TA systems. Our bioinformatics analysis, however, reveals a limited occurrence of multiple TA instances within clades and that cross-talk between clades is disfavored. These challenges the assumption of cooperativity in TA action, possibly explaining the absence of spontaneous persister generation in kacAT.

Keywords: Type II toxin-antitoxin systems; antibiotic persistence; systems biology; non-lienar dynamics; gene expression regulation; bioinformatics;

Acknowledgments: This work was supported by the Science and Technology Commission of Shanghai Municipality (grants no. 19430750600 and 19JC1413000), the National Natural Science Foundation of China (grant no. 32070572), the Medical Excellence Award funded by the Creative Research Development Grant from the First Affiliated Hospital of Guangxi Medical University (grant no. XK2019025), and the Science Fund of the Republic of Serbia (grant no. 7750294, q-bioBDS).

Poster presentation

Methodology, performance and retrainability survey of intrinsic disorder predictors

Nevena Ćirić[1*], Jovana Kovačević[1]

[1]Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade, Serbia
nevena_ciric@matf.bg.ac.rs

Intrinsically disordered proteins and regions are widely distributed within most proteomes. Recent studies show that they are associated with many essential biological processes and a broad range of human diseases. Given the prevalence of disordered proteins and the growing acknowledgement of their functional relevance, considerable effort has been made by the bioinformatics community to provide computational tools to predict protein disorder. To date, based on various characteristics of protein disorder, along with variety of diverse computational approaches, numerous disorder predictors have been developed.
Over the past decade several review papers examining intrinsic disorder predictors have been published. All these papers have played a significant role in stimulating and greatly facilitating the development of this actively growing field by pinpointing the potential room for improvement. Inspired by these, in this work we aim to integrate the relevant information regarding the existing intrinsic disorder predictors from the corresponding research papers in a novel review, including latest prediction tools. In addition, for each disorder predictor, we examined the possibility of their retraining using different datasets. Here, we present an overview of 23 protein disorder prediction methods, including the thorough analysis of their advantages and weaknesses which derive from their different computational approaches. Regarding this, we precisely describe the methodology used for building the models and categorize them by different classification schemes. The performance of these models is presented by their scores from the most recent CAID competition. Additional contribution of this work is the models' retraining availability analysis. We describe in detail the predictors' implementation source code (if available) and propose a way around to overcome the obstacles with retraining procedure (if possible). This insight might be very useful, since older models were trained on significantly smaller datasets compared to the newer ones, due to the increase in the number of experimentally annotated disorder proteins with time. With respect to this, we discuss in detail the possibility of retraining the models on a different (bigger, novel) dataset in order to perform full-scale comparison of their expression power in delineating disorder in proteins.

Keywords: intrinsic disorder, predictors, review, categorization, retraining availability

Evaluating *ND1* and *Cytb* mitochondrial genes as markers for diversity analysis of protected White-tailed eagle species from Serbia

Slobodan Davidovic[1]*, Milica Stanković[1], Pavle Erić[1], Katarina Erić[1], Aleksandra Patenković[1] and Marija Tanasković[1]

[1]Department of Genetics of Populations and Ecogenotoxicology, Institute for Biological Research "Siniša Stanković"- National Institute of the Republic of Serbia, University of Belgrade, Bulevar Despota Stefana 142, 11060 Belgrade, Serbia.
slobodan.davidovic@ibiss.bg.ac.rs

White-tailed eagle is the biggest bird of prey in Central and Southeast Europe. In Serbia it inhabits the Vojvodina province and the valleys of Danube, Sava, Tisa and Tamiš. Anthropogenic pressure on its habitats in Europe caused a decline in its numbers, but due to the strict laws protecting both species and its habitats, birds' numbers are now steady and increasing. In Serbia, as a strictly protected species it is a subject of different conservation programs. The available genetic data for this population are scarce and it is necessary to assess its genetic diversity to improve the existing conservation efforts. *ND1* and *Cytb* mitochondrial genes can be used to estimate the populations' adaptation to different environmental conditions and their variability can potentially be used to evaluate differentiation between populations.

To assess the genetic diversity of White-tailed eagle in Serbia we used mitochondrial *ND1* and *Cytb* nucleotide sequences from 40 unrelated birds collected in nests. *ND1* and *Cytb* nucleotide sequences variability was evaluated using standard parameters of genetic diversity (PGD). Acquired values were compared with the available data for the variability of the *D-loop* region which showed that combined *ND1/Cytb* nucleotide sequences PGD provide comparable results. Using publicly available sequences we reconstructed haplotype networks for *ND1*, *Cytb*, *ND1/Cytb* and *D-loop* which further showed the applicability of ND1/*Cytb* in population genetics analyses. Phylogeny reconstructed using combined *ND1/Cytb* sequences identified two branches in Serbian white-tailed eagles. Although the majority of substitutions were nonsynonymous, no selective pressure was detected.

Our data suggest that combined *ND1/Cytb* sequence variability provides sufficient information to be used for population comparison, population differentiation analyses and phylogeny reconstruction, but also gives a tool to potentially identify adaptations to different environmental conditions.

Keywords: sequencing, population genetics, genetic markers, genetic diversity, protected species

Acknowledgement: We want to thank Ištvan Ham for providing the white-tailed eagle's feathers

Poster presentation

Analysis of nucleotide sequence repeats in coronaviruses

S. Kapunac[1]*, S. Malkov[1], M. Beljanski[1], G. Pavlović Lažetić[1],
B. Stojanović[2], M. Maljković[1], A. Veljković[1] , N. Mitić[1]

[1] Faculty of Mathematics, University of Belgrade,
Studentski trg 16, 11000 Belgrade, Serbia
[2]Mathematical Institute SASA,
Knez Mihaila 36, 11000 Belgrade, Serbia
stefan.kapunac@matf.bg.ac.rs

Repeats in nucleotide sequences are connected with various genome characteristics. RNA secondary structures are related to repeats at the primary structure level. Four different types of nucleotide repeats may be identified: direct non-complementary, direct complementary, inverse non-complementary and inverse complementary. Reverse complementary tandem repeats, for example, may form hairpin secondary structures, while reverse non-complementary may be recognized by proteins. On the other side, direct complementary and/or non-complementary repeats may be reflected in protein sequence repeats, if found in the same reading frame, within the protein-coding sequence. Here we analyzed (determined and compared) all four types of nucleotide repeats in referent sequences of SARS-CoV-1, SARS-CoV-2 and MERS-COV viruses. In addition to the complete repeat set, we analyze different repeat subsets: repeats with the left component within the 5' end, repeats with the right component within the 3' end, and repeats with at least one component within the surface glycoprotein coding sequence.
We found significant differences in repeat sets corresponding to analyzed sequences in all analyzed repeat sets. In this moment we can only speculate what are the real consequences of the discovered differences.

Keywords: SARS-COV, MERS-COV, SARS-COV-2, nucleotide repeats

Genomic Surveillance and Phylogenetic Analysis of SARS-CoV-2 Variants in Serbia: Insights into Evolutionary Dynamics and Genetic Diversity

Mirjana Novkovic*[1], Bojana Banovic Djeri[1],
Sasa Todorovic[1], and Valentina Djordjevic[1]

[1]Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Vojvode Stepe 444a, 11000 Belgrade, Serbia
mirjananovkovic@imgge.bg.ac.rs

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a global pandemic, resulting in significant morbidity and mortality worldwide. Understanding the evolutionary dynamics and genetic diversity of the virus were crucial for virus control and management strategies. With that aim we conducted genomic surveillance and phylogenetic analysis of SARS-CoV-2 variants in Serbia, spanning from March 2020 to the end of January 2023.

Sequencing was conducted using three different platforms: Oxford Nanopore, Ion Torrent AmpliSeq and BGISEQ-500. Consensus sequences obtained using platforms respective software were deposited in the GISAID database. In this study 2109 good-quality sequences were included (doi.10.55876/gis8.230411qh). Pangolin and Nextclade software were utilized for clade, lineage and variant determination, while sequence alignment and construction of the phylogenetic tree was performed using Nextstrain web-based application.

Variant analysis revealed over 125,000 mutations across the 2109 sequences, of which 38% occurred in the S protein encoding gene. The most common mutations involved intragenic single nucleotide variants (88%), followed by intragenic deletions (5%). All sequences were assigned to following 16 clades: 20A, 20B, 20C, 20D, 20E, 20G, 20I, 21J, 21K, 21L, 22A, 22B, 22C, 22D, 22E, and 22F.

Temporal analysis of the variants in Serbia revealed that the Alpha variant was predominant during 2020 and the first three months of 2021. The Delta variant emerged in June 2021, dominating until the end of December 2021, when Omicron variant was detected for the first time, overtaking the dominance for the remaining surveillance period. Notably, the Gamma and Epsilon variants were not detected in the analyzed samples.

Phylogenetic analysis demonstrated that the SARS-CoV-2 variants circulating in Serbia were largely comparable to the variants found in Europe. However, a slight delay in their emergence was observed, potentially attributed to a lower travel rate during that period and a decreased frequency of sequencing in certain months.

Keywords: Sars-CoV-2 variants, Next-generation sequencing, phylogeny, surveillance, Serbia

Acknowledgement: We would like to thank all researchers involved in testing and collecting Sars-Co-2 samples in Huo Yan National Laboratories for molecular detection of infectious agents (Clinical Center of Serbia in Belgrade and Clinical Center of Nis), Clinical Center of Vojvodina, Institute of Virology Vaccines and Sera "Torlak", Veterinary Specialist Institute "Kraljevo", Scientific Veterinary Institute "Novi Sad", and The Directorate for National Reference Laboratories. Further we would like to thank all the researchers who were involved in sequencing, analyzing and submitting data to GISAID from the Institute of Molecular Genetics and Genetic Engineering University of Belgrade (UB), Faculty of Medicine UB, Faculty of Biology UB, and especially to the Veterinary Specialist Institute Kraljevo who were the first to start SARS-CoV-2 sequencing in Serbia in the very beginning of pandemic.

Poster presentation

Deciphering the reward-related impulsivity domains in rats: The big data study of historical control

Jovana Aranđelović[1]*, Kristina Mirković[1], Jana Kojić[1], Miroslav Savić[1]

[1] Faculty of pharmacy, University of Belgrade,
Vojvode Stepe 450, Belgrade, Serbia
jarandjelovic@pharmacy.bg.ac.rs

Impulsivity is a lack of ability to control own impulses, and encompasses many subdomains. The variable-delay-to-signal (VDS) paradigm is behavioral procedure for assessing motor impulsivity and delay intolerance in rats, but it was unclear whether all parameters contributed to these domains. Therefore, the aim of this study was to uncover the relationship between impulsivity parameters in a large cohort.

VDS adapted to a touchscreen environment was used to assess impulsivity in adult Sprague-Dawley rats. After 1 week of training, animals were tested in a 3-stage testing protocol. The first stage included 20 trials with 6s inter-trial interval (ITI6si) that suggested motor impulsivity. The second stage, with 60 randomly distributed trials of ITI9s or 15s, was interpreted as delay intolerance, whereas for the last stage (ITI6sf), which is similar to the first stage, it was unclear to which type of impulsivity it was associated. Principal component analysis (PCA) was used to determine the different behavioral domains. The results of 132 controls from 11 independent VDS experiments were analyzed. Based on the cumulative variance explained, scree plot, and eigenvalues, the main components were extracted whereby varimax rotation was used on factor loadings to extract the components. PCA with varimax rotation was performed in R studio.

PCA revealed that 96.45% of the variance could be explained by 3 principal components (PCs). After varimax rotation, loadings for ITI9s and ITI15s were 0.8189 and 0.9419, respectively, for rotated PC1 (RC1), loading for ITI6sf was 0.9482 for RC2, and loading for 6si was 0.9183 for RC3.

In the VDS paradigm, 3 different impulsivity domains could be determined. In addition to motor impulsivity and delay intolerance, it is suggested that reflection impulsivity can also be assessed as learning-based impulsivity.

Keywords: principal component analysis, biostatistics, rat behavior, impulsivity

Computer analysis of glioma gene network structure

Iarema P.O.[1], Turkina V.A.[1], Mayorova A.A.[1], Orlov Y.L.[1*]

[1]I.M. Sechenov First Moscow State Medical University (Sechenov University)
y.orlov@sechenov.ru

Computer analysis of disease susceptibility genes using online bioinformatics tools and open databases allows the identification of potential target genes for therapy. In the course of this study we reconstructed the gene network for genes associated with glioma. The relevance of the work is due to the fact that gliomas are the most common primary brain tumors. Gliomas originate from glial cells that support and protect nerve cells in the brain and spinal cord. Despite surgical removal, gliomas are still prone to recurrence because they grow rapidly in the brain, are resistant to chemotherapy, and are very aggressive (Byun Y.H. et al, 2022).

The task was to collect a list of glioma genes, analyze gene ontologies, reconstruct the gene network, and analyze the spatial structures of the associated proteins.

The following online bioinformatics tools were used: STRING-DB (https://string-db.org/) for gene network construction, MalaCards (https://www.malacards.org/), OMIM database (https://omim.org/). The search was performed using the keyword "glioma". AlphaFold (https://alphafold.ebi.ac.uk/), PDB (https://www.rcsb.org/) resources were used to model and visualize 3D protein structures. PANTER (http://www.pantherdb.org/) and DAVID (https://david.ncifcrf.gov/summary.jsp) resources were used to analyze gene ontologies. The list of genes for analysis consisted of 176 genes.

The most significant categories for glioma genes according to DAVID are: binding of identical proteins, negative regulation of biological processes, regulation of programmed cell death, regulation of cell death, and cell population proliferation.

The gene network was reconstructed using the STRING-DB resource (https://string-db.org/). MicroRNA genes were not recognized by the program. The graph included 150 genes. The study of the gene network structure showed high connectivity of genes within certain clusters. The EGFR and TP53 genes, which are known and well-studied oncogenes, had the greatest number of connections, as well as STAT3, KRAS, PIK3CA, IDH1, KDR. Construction of the glioma gene network showed that some elements of the graph are sufficiently linked, while others are only partially linked so that the search for target proteins for glioma treatment can be facilitated.

Three-dimensional structures of KRAS and PIK3CA proteins were constructed using AlphaFold software (https://alphafold.ebi.ac.uk/). PAE viewer (http://www.subtiwiki.uni-goettingen.de/v4/paeViewerDemo) was used to check the validity of the predicted protein structure. The structure of KRAS protein was found to be similar to that of 7ROV protein obtained from PDB (https://www.rcsb.org/) and the structure of PIK3CA protein was found to be similar to that of 4YKN protein.

Poster presentation

Genome-wide association analysis for severe COVID-19 in Serbian population

Marko Zecevic[1,2], Nikola Kotur[1*], Bojan Ristivojevic[1], Vladimir Gasic[1], Branka Zukic[1], Sonja Pavlovic[1] and Biljana Stankovic[1]

[1]Laboratory for Molecular Biomedicine, Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia,
[2]Seven Bridges, Boston, MA, United States
nikola.kotur@imgge.bg.ac.rs

Host genetics, an important contributor to the COVID-19 clinical susceptibility and severity, currently is the focus of multiple genome-wide association studies (GWAS) in populations affected by the pandemic. This is the first study from Serbia that performed a GWAS of COVID-19 outcomes to identify genetic risk markers of disease severity. A group of 128 hospitalized COVID-19 patients from the Serbian population was enrolled in the study. We conducted a GWAS comparing (1) patients with pneumonia (n = 80) against patients without pneumonia (n = 48), and (2) severe (n = 34) against mild disease (n = 48) patients, using a genotyping array followed by imputation of missing genotypes. We have detected a significant signal associated with COVID-19 related pneumonia at locus 13q21.33, with a peak residing upstream of the gene KLHL1 (p = 1.91 × 10–8). Our study also replicated a previously reported COVID-19 risk locus at 3p21.31, identifying lead variants in SACM1L and LZTFL1 genes suggestively associated with pneumonia (p = 7.54 × 10–6) and severe COVID-19 (p = 6.88 × 10–7), respectively. Suggestive association with COVID-19 pneumonia has also been observed at chromosomes 5p15.33 (IRX, NDUFS6,MRPL36, p = 2.81 × 10–6), 5q11.2 (ESM1, p = 6.59 × 10–6), and 9p23 (TYRP1,LURAP1L, p = 8.69 × 10–6). The genes located in or near the risk loci are expressed in neural or lung tissues, and have been previously associated with respiratory diseases such as asthma and COVID-19 or reported as differentially expressed in COVID-19 gene expression profiling studies. Our results revealed novel risk loci for pneumonia and severe COVID-19 disease which could contribute to a better understanding of the COVID-19 host genetics in different populations.

Keywords: GWAS, SARS-CoV-2, genetic markers, pneumonia, severe disease

Acknowledgement: Genotyping of the samples was supported by COVID-19 Host Genetics Initiative. It was performed by the Human Genomics Facility of the Genetic Laboratory of the Department of Internal Medicine at Erasmus MC. Computational resources were provided by The Cancer Genomics Cloud, powered by Seven Bridges, a component of the NCI Cancer Research Data Commons (datacommons.cancer.gov), funded in whole or in part with Federal funds from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN261201400008C and ID/IQ Agreement No. 17X146 under Contract No. HHSN261201500003I.

Impact of different mapping tools on detection of small RNAs in bacterial outer membrane vesicles

Bojana Banović Đeri*[1], Sofija Nešić[1], Ana Pantelić[1], Jelena Samardžić[1], Dragana Nikolić[1]

[1] University of Belgrade, Institute of Molecular Genetics and Genetic Engineering, Laboratory for Plant Molecular Biology, Vojvode Stepe 444a, Belgrade, Serbia
bojanabanovic@imgge.bg.ac.rs

Bacterial small RNAs (sRNAs) represent a highly diverse RNA class ranging from 8 to 200 nucleotides in length, originating from the bacterial chromosome, plasmids or phages. After syntheses sRNAs can remain inside the bacterial cell, be secreted or packed into outer membrane vesicles (OMV), enabling various intra- and inter-kingdom interactions. Different sRNAs biotypes display differences in structure, mechanism of action and level of regulation (i.e. transcription, translation, mRNA stability, etc.), but could be broadly grouped in: *trans*-acting sRNAs (bind to target mRNAs) and *cis*-encoded sRNAs (or antisense RNA that may interact not only with mRNAs, but also with proteins and DNA).
Even though the advancement of high-throughput sequencing technology led to a burst of knowledge on small RNAs complexity and diversity, there are still specific challenges related to sRNA-seq data analysis that need to be resolved. Two main challenges, associated to short length of many bacterial sRNA biotypes, are: (i) to discriminate between functional sRNAs synthesized by bacterial cell and degradation fragments produced by sample preparation and (ii) to detect functional sRNAs displaying sequence variation. While loss of very small sized sRNAs could easily be overcome by cutting-off only the specific adapter sequences that were used in sRNA library preparation, providing a proper mapping still remains a strenuous task.
The aim of this study was to test five different mapping tools that are widely used in NGS data analysis (bbmap, bowtie2, bwa, minimap2 and segemehl) for their performances in mapping of bacterial OMV sRNA-seq data to bacterial reference genome. For this test publicly available NCBI sRNA-seq dataset from OMVs of *Aliivibrio fischeri* (PRJNA629425) was used, as it contained sRNAs of different length and biotype and because *A.fischeri* reference genome and annotation were available (PRJNA12986). We evaluated five mappers using alignment and assignment rates as well as computational time. Alignment rate was calculated as the ratio of aligned and input reads, while the assignment rate was calculated as the ratio of assigned and aligned reads. Finally, totals of detected sRNAs biotypes were compared between different mappers. The statistical analysis was performed in R (version 4.3.0) and performance metrics are discussed.

Keywords: small RNAs, outer membrane vesicles, mapping

Acknowledgements: This work was funded by the project "ExplOMV", ID 7744906, Program IDEAS, Science Fund of the Republic of Serbia

Poster presentation

In silico pre-selection of β-glucosidase gene for heterologous recombinant expression

Marija Atanaskovic*[1], Ivana Moric[1], Milos Rokic[1], Lidija Senerovic[1]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade 152, Serbia
matanaskovic@imgge.bg.ac.rs

Biofilms are ubiquitous in nature, and the food industry is vulnerable to the risks posed by biofilm formation. Not only do they interfere with the food production process, but they also pose a public health threat. However, complete elimination of biofilms on food and food contact surfaces cannot be achieved by conventional methods (cleaning and disinfection) alone. New biofilm control strategies must be developed to prevent its formation and/or persistence. Novel approaches may be based on enzymes that depolymerize components of the biofilm matrix, making bacterial cells accessible to antimicrobial agents.
Environmental microorganisms are an inexhaustible source of new enzymes. In *Salmonella Enteritidis* and *Escherichia coli*, known foodborne pathogens, cellulose is an important component of the biofilm matrix, so our isolates from untapped environments were tested for cellulolytic activity. Of the more than 70 isolates examined, isolate BG28 was selected as the most promising. Its genome was sequenced, annotated, and it was identified as Gram-positive *Microbacterium* sp. Genome mining revealed the presence of four complete genes for different β-glucosidases, one of three enzyme types of cellulase complexes. To select the best candidate for heterologous expression DeepTMHMM, ProtParam, and SoluProt were used to predict the presence/absence of signal peptide and transmembrane domains, instability index, aliphatic index, hydrophilicity, and soluble expression in *E. coli*. Based on the prediction results, the gene annotated as β-glucosidase B was selected for recombinant expression. In addition, I-TASSER was used to model the tertiary structure of the selected enzyme.
The β-glucosidase B was recombinantly expressed, purified, and tested for its anti-biofilm activity. It was active and showed a 50% inhibitory effect on *S. Enteritidis* and *E. coli* biofilm formation at a concentration of 100 μg/ml. To further evaluate this in silico approach in the preselection of candidate enzymes for recombinant expression and purification, we will use it to identify other enzymes of the cellulase complex.

Keywords: β-glucosidase, bioinformatics tools, enzyme, biofilm

Acknowledgement: This study has been funded by the Ministry of Education, Science and Technological Development, Serbia (grant number 451-03-68/2022-14/200042).

Supervised Machine Learning Approach for Prediction of Occult Lymph Node Metastasis in T1-T2 Papillary Thyroid Carcinoma

Marina Popović Krneta[1]*, Nemanja Krajčinović[2], Zoran Bukumirić[3], and Miljana Tanić[4,5]

[1] Department of nuclear medicine, Institute for oncology and radiology of Serbia, Pasterova 14, 11000 Belgrade, Serbia
[2] Faculty of Technical Sciences, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia
[3] Institute of Medical Statistics and Informatics, Faculty of Medicine, University of Belgrade, Dr Subotića 8, 11000 Belgrade, Serbia
[4] Department of Experimental Oncology, Institute for Oncology and Radiology of Serbia, Pasterova 14, 11000 Belgrade, Serbia
[5] UCL Cancer Institute, 72 Huntley St London WC1E 6DD, United Kingdom
marina.popovic1989@gmail.com

This study aimed to assess and compare four machine learning (ML) based classifiers in predicting occult cervical lymph node metastasis (LNM) in clinically node-negative (cN0), T1-T2 papillary thyroid carcinoma (PTC) patients.
The study cohort included 288 PTC patients who underwent total thyroidectomy and prophylactic central neck dissection with sentinel lymph node biopsy performed for lateral LNM identification. The classifiers, namely k-Nearest Neighbor (k-NN), Support Vector Machines, Decision Tree, and Logistic Regression were developed using patients' demographic and clinicopathological variables. Evaluation metrics such as area under the receiver operating characteristic curve (AUC), sensitivity, specificity, positive and negative predictive values (PPV and NPV), accuracy, and F1 and F2 scores were utilized for model comparison.
The final ML classifier was selected based on achieving the highest specificity and the lowest degree of overfitting while maintaining a sensitivity of 95%. Among the evaluated models, the k-NN emerged as the best-performing, demonstrating an AUC of 0.72. The sensitivity, specificity, PPV, NPV, F1, and F2 scores were 98%, 27%, 56%, 93%, 72%, and 85%, respectively Furthermore, a web application was developed allowing users to predict the potential of cervical LNM and explore possibilities for further model development.
The k-NN classifier incorporating patients' clinicopathological information shows potential in predicting LNM. Improved prediction models are necessary to identify patients at higher risk of LNM, guiding appropriate postsurgical treatment for high-risk individuals while minimizing unnecessary interventions for low-risk patients.

Keywords: machine learning, papillary thyroid carcinoma, lymph node metastasis

Acknowledgement: This research was supported by the Serbian Ministry of Science, Innovation and Technological development (451-03-47/2023-01/200043).

Poster presentation

Determinants of CRISPR array non-canonical adaptation mechanism

Marko Tumbas[1]* and Marko Đorđević[1]

[1] Quantitative Biology Group, Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
marko.tumbas@bio.bg.ac.rs

CRISPR-cas systems are incredibly diverse and currently are classified in six major types and over 30 subtypes. Apart from their role in adaptive immunity it has been shown that some of the CRISPR-cas subtypes are also involved in host gene regulation and even in collateral damage leading to bacteriostatic or lethal outcomes for the host. CRISPR array spacers direct and influence canonical and non-canonical functions of the CRISPR-cas system together with subtype Cas proteins. Better understanding of spacer adaptation mechanisms is crucial for uncovering intricacies of evolutionary arms race between prokaryotes and phages.

Here we present large-scale analysis of CRISPR array spacers originating from 31845 complete bacterial genomes. All bacterial and 16388 viral genomes were retrieved using NCBI datasets API. CRISPRidentify and CRISPRcasIdentifier tools were used for CRISPR array, Cas genes detection and subtyping. Viral genomes were mapped to their hosts using the latest version of the Virus-Host DB. Mapping was performed on the genus level of the hosts phylogenetic tree. Gumbel extreme value distribution was used to determine statistical significance of each spacer Smith-Waterman alignment score.

Differences in melting energy and GC content between identified spacers, origin bacterial genomes and infecting bacteriophages were explored for different CRISPR-cas subtypes and for different bacterial genera. Spacers from the extremes of the GC content distribution were aligned to the origin bacterial and infecting phage genomes in order to determine their origin.

GC content of the spacers was lesser than the GC content of the source bacterial genome but greater than infecting viral genome. This observation aligns with the hypothesis that the majority of CRISPR spacers were adapted from the bacteriophage genomes and serve canonical function. Alignments of the spacers from GC rich distribution tail have shown their preferential targeting of host genomes which further supports the hypothesis that GC rich spacers originated from the bacterial genome and have non-canonical function.

Keywords: CRISPR-cas, melting energy, extreme value distribution

Data mining for long-non coding RNAs deregulated in colon cancer through analysis of Gene Expression Omnibus database

Iva Pruner[1]* and Aleksandra Nikolic[1]

[1] Institute of Molecular Genetics and Genetic Engineering,University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
iva@imgge.bg.ac.rs

Colorectal cancer (CRC) is one of the most commonly diagnosed cancers worldwide. Lack of specific CRC symptoms is a challenge for clinicians, as the symptoms overlap with other non-cancerous diseases, leading to 20-25% of newly diagnosed CRC patients already having liver metastasis. Thus, discovering reliable early-disease biomarkers is of high importance. Non-coding RNAs (ncRNAs) have been demonstrated to be involved in CRC development and progression. Long non-coding RNAs (lncRNAs) can interact with RNA, DNA and proteins, forming complexes that are involved in regulation of gene expression via multiple mechanisms, affecting every stage of colon carcinogenesis and making them top candidates for novel biomarker discovery.
The aim of our study was to conduct data mining of Gene Expression Omnibus (GEO) database by using "colon cancer" and "ncRNA" keywords, and identify differentially expressed lnRNAs present in different GEO datasets.
GEO database which collects submitted high-throughput gene expression data was queried for all datasets that studied colon cancer and ncRNA. Over 60 datasets were manually inspected in order to identify those where analysis of colon and normal tissue originating from the same patient was done. Each dataset was analyzed by GEO2R software to discover differentially expressed lncRNAs. LncRNAs were considered significant if they appeared in more than one GEO dataset. Parts of lncRNAs sequences available in GEO2R analysis results were run through BLAST in order to identify full length lncRNAs.
Five GEO datasets matched our criteria. We discovered 12 sequences that appeared in more than one dataset and we identified them through BLAST analysis. Six sequences originated from lncRNAs (RYR3 divergent transcript, long intergenic non-protein coding RNA 595, TOX divergent transcript, FLVCR2 antisense RNA 1, LHRI_LNC744.1 lncRNA gene, and ELFN1 antisense RNA 1), while six sequences represented partial sequences of various mRNAs. Four lncRNAs were down-regulated in colon cancer; one was up-regulated, while one showed different expression patterns in different GEO datasets.
In this study, we have identified six lncRNAs that have potential significance for colorectal cancer etiology and will be a subject of further *in silico* and *in vitro* study.

Keywords: long non-coding RNA, colorectal cancer, data mining, GEO database

Poster presentation

Efficient bioinformatics workflow for *de novo* transcriptome assembly of *Pelargonium zonale*

Dejana Milić*, Ana Pantelić, Jelena Samardžić, Bojana Banović Đeri, Marija Vidović

University of Belgrade, Institute of Molecular Genetics and Genetic Engineering, Laboratory for Plant Molecular Biology, Vojvode Stepe 444a, Belgrade, Serbia
dmilic@imgge.bg.ac.rs

Variegated *Pelargonium zonale* is a widely cultivated ornamental plant characterized by green, photosynthetically active tissue (GL) and white, non-photosynthetic tissue (WL). The aim of this study was to investigate the transcriptomic differences between these two tissue types.
We performed RNA-seq analysis of GL and WL on Illumina HiSeq 2500 platform. The raw reads were processed using in-house scripts to remove low-quality reads, adapter sequences, poly-N sequences, and contaminants. High-quality clean reads were subjected to de novo transcriptome assembly using Trinity (min_kmer_cov = 2, min_glue = 2). The redundancy was removed and longest transcripts per cluster were selected as unigenes.
Gene expression levels were estimated using RSEM by mapping clean data back to the assembled transcriptome (Bowtie2 with mismatch = 0). Differential expression analysis between GL and WL (three biological replicates per each) was performed with DESeq2 R package (p values adjusted according to Benjamini and Hochberg for controlling False Discovery Rate). Genes with abs (log2 FC) ≥ 2 and adjusted p value < 0.05 were assigned as statistically significant differentially expressed. Functional enrichment analysis was performed using GOseq R package and KOBAS software (corrected p < 0.05).
We annotated 85,374 unigenes (61.17%), providing a valuable resource for future functional genomics studies. Out of 8896 gene clusters that were statistically significantly differentially expressed between the green and white leaf tissues (p value < 0.05 and abs(log2 fold change) ≥ 2), 5585 were upregulated in the WL, while 3311 were upregulated in the GL. These findings shed light on the transcriptomic differences between the two leaf tissue types in *P. zonale* and provide a foundation for further research on the functional significance of these differences. Also, this study demonstrated utility of the Trinity pipeline for *de novo* transcriptomic analysis of organism whose genomes are yet not sequenced.

Keywords: *de novo* transcriptomic assembly, variegated plants, *Pelargonium zonale*, Trinity software

Acknowledgements: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-47/2023-01/ 200042) and Bilateral project (no. 451-03-01963/2017-09/09).

Application of principal component analysis (PCA) and analytical hierarchy process (AHP) in analysis of articulatory characteristics of phonemes of children with 22q11.2 Deletion Syndrome

Danijela Drakulic[1]*, Marijana Rakonjac[2,3], Goran Cuturilo[4,5], Natasa Kovacevic-Grujicic[1], Jelena Kusic-Tisma[1], Ivana Moric[1], Branka Zukic[1], and Milena Stevanovic[1, 6,7]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade 152, Serbia
[2] Institute for Experimental Phonetics and Speech Pathology, Jovanova 35, 11000 Belgrade, Serbia
[3] Speech and language pathology center "Logogen"
[4] University Children's Hospital, Tirsova 10, 11000 Belgrade, Serbia
[5] Faculty of Medicine, University of Belgrade, Dr Subotica 8, 11000 Belgrade, Serbia
[6] Faculty of Biology, University of Belgrade, Studentski trg 16, 11000 Belgrade, Serbia
[7] Serbian Academy of Sciences and Arts, Kneza Mihaila 35, 11000 Belgrade, Serbia
danijeladrakulic@imgge.bg.ac.rs

22q11.2 deletion syndrome (22q11.2DS is caused by 22q11.2 microdeletion, one of the strongest known risk factors for development of neurodevelopmental disorders. About 70% patients with 22q11.2DS have speech and language impairments. In the literature, there is no data about articulatory characteristics of phonemes of children with 22q11.2DS, monolingual native speakers of South Slavic languages. Here we, by applying Global Articulation Test, analyzed articulatory characteristics of phonemes of children with 22q11.2DS, monolingual native speakers of the Serbian language (group E1), children with a phenotype resembling 22q11.2DS but without the microdeletion (group E2), children with non-syndromic congenital heart malformations (since children with these malformations may exhibit a speech and language impairments) (group E3) and their peers with typical speech-sound development (group C). Results of PCA indicated that the groups can be distinguished based on the pronunciation of phonemes, and that the pronunciation of the phonemes "Č ⟨tʃ⟩", "Dž ⟨dʒ⟩", "Š ⟨ʃ⟩", "Ž ⟨ʒ⟩", "R", and "Lj ⟨ʎ⟩" contributes the most to the variability between the groups. Results of AHP revealed that the pronunciation of the phonemes "Č ⟨tʃ⟩", "Dž ⟨dʒ⟩", "Š ⟨ʃ⟩", "Ž ⟨ʒ⟩", "R", and "Lj ⟨ʎ⟩" was rated the worst in the group E1. In conclusion, obtained results indicate that the presence of 22q11.2 microdeletion influences articulation skills of carriers.

Keywords: PCA, AHP, articulation, 22q11.2DS

Acknowledgement: This research was funded by European Union's Horizon Europe programme (Grant Agreement Number 101060201 (STREAMLINE)), Ministry of Science, Technological Development and Innovation of the Republic of Serbia (grant number 451-03-47/2023-01/200042) and the Serbian Academy of Sciences and Arts (Grant number F-172).

Poster presentation

Integrated relational database of human protein-protein interactions

Bojana Jošić[1]*, Jovana Kovačević[1], Vladimir Perović[2], Nevena Veljković[2]

[1] Faculty of Mathematics, University of Belgrade,
Studentski trg 16, Belgrade, Serbia
[2] Institute of Nuclear Sciences Vinča, University of Belgrade,
Mike Petrovića Alasa 12-14, Belgrade, Serbia
mr17128@matf.bg.ac.rs

Protein-protein interactions' data are stored in various publicly available databases of different types and formats. In this work, a new database for protein-protein interactions is created by integrating data from multiple existing databases. This task is not trivial since different databases use distinct gene or protein identifiers for protein annotation. Additionally, they use different methods to determine interaction scores, and the interactions are obtained through diverse experimental or predictive methods. As a result, two databases may store different data about the same interaction.

To integrate data from various databases, namely *BioGRID*, *STRING*, *HIPPIE*, *IntAct*, and *Reactome*, into a single PPI database, the following process is undertaken. Initially, data is downloaded from these databases in the MITAB format, encompassing all pertinent interaction information such as protein identifiers, publication sources and other. In order to obtain unique protein identifiers in all PPIs in the database, the *UniProt ID mapping* tool was used to determine *UniProt IDs*. Next, since scoring systems differ among databases, for every interaction a new score is calculated using *MISCORE* tool as an additional metrics unique for all the PPIs in the database. The resulting database contains tens of millions of human PPIs from five different sources.

Keywords: bioinformatics, protein-protein interaction, database, computer science

Mining for the data about glycosylation in the bovines-the analysis of the recently published studies

Anđelo Beletić[1]*, Ivana Duvnjak Orešković[1], Tea Pribić[1], and Gordan Lauc[1,2]

[1] Genos Ltd, Glycoscience Research Laboratory,
Borongajska cesta 83H, 10000 Zagreb, Croatia
[2] Faculty of Pharmacy and Biochemistry, University of Zagreb,
Ante Kovačića 1, 10000 Zagreb
abeletic@genos.hr

Deciphering the glycosylation patterns and mechanisms in bovines (*Bos taurus*) holds the potential for improvement regarding reproduction, herd health management, and the quality and safety of milk and meat products. The PubMed database was searched for "glycosylation" and "B. taurus" using the following filters: full text available, the publication date of five years, and the preprints excluded. The search retrieved 244 results, and after the content analysis by the authors, 88 remained relevant. All publications were Research Articles except one Review. The assessment of the glycan profile composition was among the aims in 34, the functional aspects in 33, and the protein glycoforms in 12 studies. Ten studies brought data about the total glycome profile of the milk, tissue, or meat sample, while the other contained glycosylation-related features of the individual protein(s). Most often, the studies used milk (25), individual proteins (23), or tissue (20 studies) as the samples. Usually, the milk was material to analyze the glycosylation of casein, immunoglobulin G, or the total glycans. The studies involving the individual proteins most frequently analyzed fetuin, and the glycosylation of submaxillary gland mucin was the target in the studies using tissue samples. These pioneer data mining results allow for the conclusion on the availability of reliable data about glycosylation in the bovines, eligible as the starting point for further scientific efforts on their continuous appending, systematization, and multidisciplinary analyses.

Keywords: data mining, glycosylation, bovines

Poster presentation

Different approaches in microRNA analysis

Barbara Jenko Bizjan[1,2]*, Bine Stančič[1], Iva Sabolić[3],
Maja Štalekar[1] and Uršula Prosenc Zmrzljak[1]

[1] BIA Separations CRO, Labena d.o.o., Ljubljana, Slovenia
[2] Clinical Institute of Special Laboratory Diagnostics, University Children's Hospital, UMC, Ljubljana, Slovenia
[3] eDNA labs, Labena d.o.o., Zagreb, Croatia
barbara.jenko@biaseparationscro.com

MicroRNA might serve as a predictive biomarker for treatment response in stem cell treatment in knee osteoarthritis. Different sample types are going to be collected to enlighten the true biological role. MicroRNA analysis necessitates diverse approaches based on the sample type. In this study, we examined microRNA profiles in plasma samples, synovial fluid, and adipose-derived fat tissue. We conducted a comparative analysis of different microRNA analysis methods to assess the data.
The first approach involved a series of steps, including adapter trimming, quality filtering, size filtering, and mapping of all reads to the human reference genome (GRCh38.p12). Subsequently, genome-mapped reads were aligned to known miRNA sequences from miRBase. Reads that did not match miRNAs were subjected to further classification using additional databases, such as RNAcentral. The second pipeline also encompassed adapter trimming, quality filtering, and size filtering. Additionally, it involved collapsing individual reads into repeat sequences, followed by alignment to the mature index of miRBase. Unaligned reads were classified as isomiRs based on their alignment to the hairpin index of miRBase.
We processed sequences from three plasma samples, three adipose fat tissue samples, and three synovial fluid samples. Although there were slight variations in microRNA read counts, the average ratio between counts was 0.92 (SD=0.29). Notably, the second pipeline yielded higher read counts compared to the first pipeline.
The results obtained from both microRNA bioinformatic pipelines demonstrated similar outcomes, suggesting that the choice of pipeline is unlikely to have a significant impact on the derived biological insights.

Keywords: bioinformatics, microRNA, sequencing

Acknowledgement: This work was performed as a part of project: "Development of advanced prediction tool for successful and optimized treatment course in pathological joint changes based on quantification of inflammatory biomarkers - PredicTest" which is co-funded by EUREKA member countries and the European Union Horizon 2020 Framework Programme

In Silico analysis and prediction of novel pharmacogenomic markers of pediatric ALL treatment

Vladimir Gašić*[1], Nikola Kotur[1], Biljana Stanković[1], Đorđe Pavlović[1], Marina Jelovac[1], Jelena Perić[1], Bojan Ristivojević[1], Sonja Pavlović[1], and Branka Zukić[1]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11000 Belgrade, Serbia
vlada.gasic42@gmail.com

Acute lymphoblastic leukemia (ALL) is the most common childhood neoplasm. Side effects of therapy occur in 75% of patients and 1-3% of patients have a lethal outcome due to treatment. More efficient treatment of pediatric ALL has been developed by avoiding drug adverse effects included in the treatment protocols. Therefore, implementation of pharmacogenomics is paramount in pediatric ALL treatment. Next generation sequencing (NGS) contributed to discovery of novel genetic markers, potential candidates for targeted therapy and predictors of efficacy and toxicity of drugs.
We aimed to discover novel potential pharmacogenomic markers in pediatric ALL.
DNA samples from bone marrow of 17 pediatric ALL patients were analyzed using the platform TruSeq Amplicon – Cancer Panel (Illumina) for somatic mutations in 48 oncogenes. DNA samples from blood of 100 individuals, using the platform TruSightOne (Illumina), were analyzed for germinative mutations. An in-house virtual panel for GC response markers was designed. Predicting the effects of novel variants was performed using the SIFT, PolyPhen-2 and PROVEAN software tools. For protein structure stability and modeling we used STRUM method and i-TASSER server.
In the NGS study of somatic mutations in pediatric ALL, 9 novel variants have been identified. Bioinformatic analysis has shown that *STK11* c.1023G>T and *ERBB2* c.2341C>T possess potential as pharmacogenomic markers, therefore, they are candidates for molecular targeted therapy. In the exome sequencing study, according to the prediction algorithms, 3 new potential markers in pharmacogenes related to GC response have been identified, *ABCB1* c.947A>G, *NCOA3* rs138733364 and *TBX21* rs14059812.
Using NGS analysis and prediction algorithms, we have detected 2 novel somatic mutations, candidates for targeted molecular therapy, as well as 3 novel germinative variants, potential pharmacogenomic markers of GC response in pediatric ALL. Pharmacogenomic profiling of each pediatric ALL patient is indispensable for new therapy approaches and it could lead to better outcomes.

Keywords: Acute lymphoblastic leukemia, Pediatric, Pharmacogenomics

Acknowledgement: This research was funded by the PharmGenHUB Project 101059870, Twinning Western Balkan call: HORIZON-WIDERA-2021-ACCESS-02

Poster presentation

Exploring Changes in Diagnoses during the COVID-19 Era: Comparative Analysis

Despina Misheva[1*], Marija Stojcheva[1], Hana Hasanicaj[1],
Ana Mladenovska[1], Jovana Dobreva[1], Mary Lucas[2], Irena Vodenska[3],
Lou Chitkushev[2], Dimitar Trajanov[1,2]

[1] Faculty of Computer Science & Engineering, Ss. Cyril and Methodius University, Rudzer Boshkovikj 16, 1000 Skopje, Macedonia
[2] CH- Boston, MA, US
[3] Administrative Sciences Department, Metropolitan College, Boston University - Boston, MA, US
despina.misheva@students.finki.ukim.mk

The healthcare sector is just one of several areas of society that have been significantly impacted by the COVID-19 pandemic. This paper aims to analyze the changes observed in the medical profession's approach to diagnosing diseases between the pre-pandemic year of 2019 and the pandemic year of 2020. By examining these shifts, we explore how medical professionals have adapted their treatment strategies, leading to modifications in diagnosis for various diseases.
Based on our visualization, shown in Figure 1, we observed that the diagnoses of ***Obstructive Sleep Apnea*** and ***End stage renal disease*** had consistent distributions in both 2019 and 2020. Also we need to mention, the count value for ***Obstructive Sleep Apnea*** was higher in 2020, whereas in 2019, the count value was higher for ***End stage renal disease***, showing their representation in each year. We can conclude that the pandemic has resulted in a marked increase in the occurrence of specific diagnoses compared to the previous year, some of them being ***acute pharyngitis-sore throat (J029)***, ***gastro-oesophageal reflux disease (K219***) and ***pure hypercholesterolemia - unspecified (E7800)***, as can be seen on Figure 1.
A notable variation can be observed when examining the months of November and December in 2020. In these months, the diagnosis ***Contact with and (suspected) exposure to other viral communicable diseases*** transitions from the third to the second position, indicating a higher occurrence of COVID-19 in December compared to November. This shift in ranking provides valuable insights into the increased prevalence of this diagnosis during the month of December. Through this analysis, we aim to examine the transformations that have taken place as a result of the pandemic, particularly in terms of the diagnosis of a specific disease, which has undergone notable changes compared to the pre-pandemic period. We highlight several significant changes that have occurred in defining diagnoses, showcasing the variations observed over the course of a year.

Keywords: COVID pandemic,
data analytics, data visualization

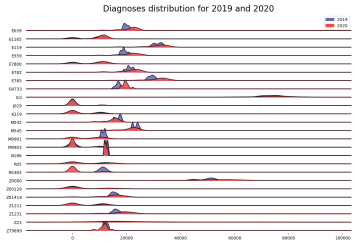


Figure 1. Diagnoses distribution for 2019 and 2020

Shotgun metagenomics reveals gut microbiota features associated with the efficacy of myeloid derived suppressor cells in the prevention of neuroinflammation

Marina Bekić[2], Nataša Ilić[2], Jelena Đokić[1], Dušan Radojević*[1], Dragana Vučević[3], Saša Vasilev[2], and Sergej Tomić[2]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
[2]Institute for the Application of Nuclear Energy, University of Belgrade, Banatska 31b, 11080 Belgrade, Serbia
[3]Medical Faculty, Military Medical Academy, Defence University, Crnotravska 17, 11000 Belgrade, Serbia
dradojevic@imgge.bg.ac.rs

Although genetic predisposition to Multiple Sclerosis (MS) may play an essential role in disease development, myeloid cell overactivation and gut microbiota dysbiosis are key contributors to MS pathogenesis. Myeloid-Derived Suppressor Cells (MDSC)s are immature myeloid cells with strong immunosuppressive functions which can be exploited in the treatment of autoimmune diseases. Considering the limited data on MDSCs application in MS therapy and their poorly studied effects on the gut microbiota, we have investigated the therapeutic potential of mice MDSC differentiated according to the standard protocol (MDSC) and modified with the addition of prostaglandin (PG) E2 (MDSC-PGE2) to ameliorate experimental autoimmune encephalomyelitis (EAE) induced with MOG35-55/CFA/PtX in C57BL/6 mice. Additionally, we analyzed the changes in gut microbiota features in control and MDSC-treated animals by using a shotgun metagenomics approach. In mice, PGE2-activated MDSC significantly inhibited the onset and clinical course of EAE. This effect correlated with increased IL-10, TGF-β, IL-4 production, and Arginase-1 level in MDSC-PGE2, as well as with reduced leukocyte infiltrates in the spinal cord. MDSC-PGE2 protective effect is also reflected in the maintenance of gut microbiota composition based on Kraken2/Bracken2 and LEfSe analysis. We observed an increase of MS-associated species *Romboutsia ilealis* in the control EAE group, while in both MDSC treatments the increase in relative abundances of *Muribaculum gordoncarteri* and *Duncaniella dubiosis*, associated with immunoregulatory properties, was observed. Microbial metabolic pathways profiling using Humann3 pipeline also reveals the increase in pathways involved in the production of potentially immunoregulatory metabolites in the MDSC-PGE2 group. In conclusion, we pointed to the significant association between the efficacy of MDSC-PGE2 treatment and gut microbiota features which can be further exploited in order to improve MDSC-based EAE therapy.

Keywords: Myeloid derived suppressor cells, gut microbiota, functional pathways, multiple sclerosis, immunoregulatory mechanisms

Acknowledgement: This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia under Contract No. 451-03-47/2023-01/200042 and No. 451-03-47/2023-01/200019, and by the Science Fund of the Republic of Serbia, PROMIS, #6062673, Nano-MDSC-Thera.

Poster presentation

Seeking an optimal variant calling pipeline for medical genetics

Yury A. Barbitoff[1]*, Alexandra Panteleeva[1], Alexander V. Predeus[1]

[1] Institute of Bioinformatics Research and Education, Belgrade, Serbia
barbitoff@bioinf.institute

Accurate and comprehensive variant discovery is extremely important for rare disease diagnostics using next-generation sequencing (NGS) methods. Over the recent years, a plethora of methods have been developed for short variant calling from NGS data, and the most recent tools extensively use machine learning algorithms for both variant discovery and filtering. In our study, we took an effort to systematically evaluate the performance of different pipelines for short variant calling in the human genome.
To perform such a systematic comparison, we collected a large dataset of both "gold standard" (provided by the Genome In A Bottle (GIAB) consortium) and in-house whole-exome sequencing (WES) and whole-genome sequencing (WGS) datasets. (a total of 20 different datasets was used). We tested all combinations of 4 popular short read aligners (BWA, Bowtie2, Isaac, and Novoalign) and 9 novel and well-established variant calling and filtering methods (Freebayes, Clair3, DeepVariant, Genome Analysis ToolKit (GATK), Octopus, Strelka2). We also used several different tools for preprocessing of short reads.
Our analysis showed negligible effects of adapter trimming on the accuracy of short variant calling. Among read aligners, Bowtie2 performed significantly worse than other tools, suggesting it should not be used for medical variant calling. For pipelines based on BWA, Isaac, and Novoalign, the accuracy of variant discovery mostly depended on the variant caller and not the read aligner. DeepVariant consistently showed the best performance and the greatest robustness compared to all other tested variant callers. We have also compared the consistency of variant calls in GIAB and non-GIAB samples. With few important caveats, best-performing tools have shown little evidence of overfitting.
Taken together, our study showed that modern strategies for NGS data analysis allow for high accuracy of genetic variant discovery within coding regions of the human genome. However, there is still a need for development of new library preparation and variant calling methods to enhance variant discovery in the challenging regions of the human genome.

Keywords: pipeline, variant calling, human genetics, medical genetics

Acknowledgement: We thank JetBrains Ltd. for providing financial support and computing resources for the project.

Groundwater and soil as a reservoir for polyurethane-degrading bacteria

Milica Ciric[1]*, Brana Pantelic[1], Vladimir Šaraba[1], and Jasmina Nikodinovic-Runic[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a
11000 Belgrade, Serbia
milica.ciric@imgge.bg.ac.rs

Plastic waste is a global environmental burden. Polyurethanes (PU), toxic and ubiquitous synthetic polymers, do not biodegrade quickly, leading to their rapid accumulation in the soil and water environments. Highly efficient PU-degrading microorganisms are rare in nature and are of fundamental importance for achieving circular plastic economy. Bacterial isolates from groundwater, originating from magmatogenic massif and Tertiary basin within metamorphic area, as well as soil isolates collected from various pristine (PS) and contaminated sites (CS), were screened using PU model compound Impranil® DLN-SD (IMP) as sole C source to identify PU-degrading isolates. Phylogenetic analysis of 16S rRNA gene sequences from IMP-degrading isolates was performed using the neighbor-joining method to observe their clustering. Thirty one of 96 isolates (32.3 %) from groundwater and 18 of 220 isolates (8.2%) from soil produced prominent IMP-clearing zones. Thirteen IMP-degrading isolates from each type of environment, belonging to 8 genera (*Pseudomonas*, *Proteus*, *Enterobacter*, *Flavobacterium*, *Serratia*, *Pantoea*, *Acinetobacter* and *Stenotrophomonas*) for groundwater and to 6 genera (*Streptomyces*, *Pseudomonas*, *Rhodococcus*, *Achromobacter*, *Bacillus* and *Paenibacillus*) for soil environment, were included in phylogenetic analysis. No clear grouping of groundwater and soil isolates was observed, indicating that isolates are too distinct. Stronger clustering was observed for groundwater compared to soil isolates. For groundwater, strongest clustering was observed for 2 isolates belonging to *Proteus* genus, 2 belonging to *Flavobacterium* and 2 to *Pseudomonas*. For soil samples, strongest clustering was observed for 3 isolates belonging to genus *Streptomyces*. There was no clear grouping within isolates from CS and PS. In the future, wider range of environmental niches should be included in screening efforts for development of biocatalytic processes for management of plastic waste. Subterranean ecosystems, which are not readily accessible for sampling and represent largely unexplored reservoir of biotechnologically relevant enzymatic activities, should also be more represented in such screenings.

Keywords: groundwater, soil, polyurethane-degrading bacteria, 16S phylogeny

Acknowledgement: This work was supported by the EU H2020 Research and Innovation Programme (grant agreement No. 870292, BioICEP) and by the Ministry of Science, Innovation and Technological Development of the Republic of Serbia (agreement No. 451-03-47/2023-01/ 200042). 16S rDNA sequences are deposited in the NCBI GeneBank database (accession numbers: OQ991477-OQ991494) for groundwater isolates and at https://doi.org/10.3390/catal13020278 for soil isolates.

Poster presentation

Developing bioinformatics pipeline for processing environmental DNA metabarcoding sequencing data

Iva Sabolić[1]*, Lucija Markulin[1], Teja Petra Muha[2],
Barbara Jenko[2], Uršula Prosenc Zmrzljak[2]

[1] eDNA Labs, Labena d.o.o., Jaruščica 7, 10000 Zagreb, Croatia
[2] BIA Separations CRO, Labena d.o.o., Teslova ulica 30, 1000 Ljubljana, Slovenia
iva.sabolic@labena.hr

Environmental DNA (eDNA) is DNA present in an environmental sample, originating from any biological material released from organisms living in that environment. This DNA can be isolated, amplified, sequenced, and analyzed in order to examine the taxonomic richness and abundance of different organism groups in the targeted environment. Methods of eDNA metabarcoding thus offer a unique opportunity to systematically streamline and scale-up regular biological assessments across many different environments of interest.
Recently, as a part of the project funded by European structural and investment funds, Labena d.o.o. company established a modern laboratory in Zagreb focused on the research and provision of services in the field of eDNA. In collaboration with the Institute Ruđer Bošković we have been working on developing tests for analysis of water quality based on the eDNA and, as part of the standardization and optimization of sample-to-results eDNA analysis process, we developed a custom bioinformatics pipeline to facilitate efficient and effective eDNA sequencing data analysis.
The pipeline was was written in Bash and utilizes several different algorithms to filter, trim, merge, denoise and classify targeted eDNA sequences. Python-based scripts which allow automatically download, filter, and format the data available on various online platforms were included in the pipeline to facilitate the curation of custom reference databases needed for taxonomic classification of targeted organism groups. User-friendly and interactive pipeline report generation, comprised of both wet- and dry-lab step-by-step sample statistics and graphical representations or the main results, is supported using Rmarkdown and Plotly and DataTables libraries. The pipeline is containerized in Docker, allowing for easier environment building and pipeline deployment.

Keywords: environmental DNA, pipeline, reference databases, containerization

Evaluation of variant calling tools for detection of SNVs in BRCA1 and BRCA2 genes in patients from the Institute of Oncology and Radiology of Serbia

Isidora Pantović[1], Katarina Živić[1], Ivana Boljević[1],
Milica Nedeljković[1], Radmila Janković[1], Miljana Tanić[1]

[1]Institute of Oncology and Radiology of Serbia, Pasterova 14,11000 Belgrade, Serbia
m1023_2022@stud.bio.bg.ac.rs

Serbia has one of the world's highest incidences and mortality rates of ovarian cancer. Germline or somatic mutations in *BRCA1* and *BRCA2* genes, such as single nucleotide variants (SNVs), indels, insertions, deletions, commonly lead to development of breast and ovary cancer. Targeted therapy with PARP inhibitors is the current standard of care for serous epithelial BRCA-mutated ovarian cancer and depends on the accurate detection of mutations in these genes.

In this study, a subset of patient specimens from Institute of Oncology and Radiology were sequenced on MiSeq Illumina sequencer, raw data were analysed bioinformatically, which included checking quality control of raw FASTQ sequences, trimming, mapping them on reference genome(hg19), target coverage quality control and variant calling. We tested various variant calling tools including Mutect2, GATK HaplotypeCaller, FreeBayes, VarDict and MuSe callers. We evaluated the relative performance- concordance rate, false positive and false negative rates between the callers for SNV/indel detection in *BRCA1* and *BRCA2* genes.

Keywords: genomics, variant calling, BRCA1 gene, BRCA2 gene, cancer, sequencing

Acknowledgement: This research was supported by the Serbian Ministry of Science, Innovation and Technological development (451-03-47/2023-01/200043).

Poster presentation

Transcriptome analysis of *Pseudomonas aeruginosa* after MhqO dioxygenase treatment

Andjela Djokic[1]*, Ivana Moric[1], Lidija Senerovic[1] and Lidija Djokic[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11000 Belgrade, Serbia
adjokic@imgge.bg.ac.rs

Pseudomonas aeruginosa is an opportunistic pathogen that can cause severe chronic infections due to its exceptional ability to form a biofilm. Regulation of biofilm formation is very sophisticated and involves multiple bacterial systems and regulatory pathways. We found an enzyme MhqO dioxygenase from *Bacillus paralicheniformis* ZP1, which was effective in the inhibition of biofilm formation and disruption of mature biofilm of *P. aeruginosa*. Our results suggest that MhqO exerts its effect at the adhesion level, preventing cells from attaching to the surface. We have also shown that the enzyme stimulates the rhamnolipids synthesis.

To elucidate the mechanism of enzyme action, we analyzed the transcriptome of the *P. aeruginosa* PAO1 strain treated with MhqO. Since cell adhesion occurs at the beginning of the stationary phase growth, the PAO1 strain was treated with MhqO for four hours, followed by total RNA isolation and cDNA synthesis. Transcriptome sequencing was performed by Illumina NovaSeq 6000 and data were analyzed by Novogene Bioinformatics Technology Co., Ltd. (Beijing, China).

Obtained data showed that 122 genes were up-regulated, 41 genes were down-regulated, and the expression of 5947 genes was not changed. Five genes whose expression was altered are directly related to biofilm formation. MhqO increased the expression of the RsmA post-transcriptional regulator in *P. aeruginosa*. Transcriptome data revealed that pili IV biosynthesis genes were up-regulated, which is in accordance with literature data that RsmA positively regulates these genes. The inhibition of cells' attachment to the surface could be explained by these results. In addition, RsmA positively regulates rhamnolipid production but negatively regulates biofilm matrix synthesis, which was supported by expression levels in the sequenced transcriptome.

Data obtained from transcriptome analysis suggest that *P. aeruginosa* treated with MhqO dioxygenase should be more sensitive to oxidative and osmotic stress, as well as to beta-lactam antibiotics. Our further investigations should confirm these effects at the phenotypic level as well.

Keywords: transcriptome, MhqO, RsmA, *Pseudomonas aeruginosa*

Acknowledgement: This study has been funded by the Ministry of Education, Science and Technological Development, Serbia (grant number 451-03-47/2023-01/ 200042).

PACSIN2 modifies miRNAs in extracellular vesicles, modulating thiopurine response

Alessia Norbedo[1]*, Marianna Lucafò[1], Carlotta Bidoli[1], Marco Gerdol[1], Metka Lenassi[2], Giuliana Decorti[1,3], Gabriele Stocco[1]

[1] University of Trieste, Department of Life Science Trieste, Italy
[2] University of Ljubljana, Faculty of Medicine, Institute of Biochemistry, VrazovTrg 2, 1000 Ljubljana, Slovenia
[3] Institute for Maternal and Child Health I.R.C.C.S. Burlo Garofolo, Trieste, Italy
alessia.norbedo@phd.units.it

Thiopurines, such as mercaptopurine, are antimetabolites, used in the treatment of acute lymphoblastic leukemia (ALL) and inflammatory bowel disease (IBD). *PACSIN2* rs2413739 is associated with gastrointestinal toxicity in children with ALL and with drug-efficacy in IBD pediatric patients. PACSIN2 is involved in vesicular trafficking and may affect the release and content of extracellular vesicles (EVs), which mediate cell communication and whose cargo modifies phenotypes of target cells. This study evaluates mechanisms associating *PACSIN2* polymorphism with interindividual variability in efficacy of thiopurines, by considering the role of PACSIN2 in sorting specific miRNA in EVs.
Effects of stable *PACSIN2* knock-down (KD) were evaluated in intestinal LS180 cells. MTT cytotoxicity assay was used to verify mercaptopurine-sensitivity. EVs, released by LS180 KD and MOCK control cells were isolated by ultracentrifuge and characterized by nanoparticle tracking analysis (NTA). EVs miRNA-sequencing was performed by Illumina Hi-seq 2000. EVs may alter drug cytotoxicity, therefore LS180 MOCK and KD cells were co-treated with mercaptopurine and EVs. Statistical analysis was performed using t-test and ANOVA.
Mercaptopurine was more cytotoxicity in LS180 KD cells (IC50 MOCK 3.23; IC50 KD 2.18 μM). No differences were observed by NTA in release of EVs between MOCK and KD cells (t-test, $p = 0.13$). *PACSIN2* KD altered intracellular and EVs expression of 6 and 24 miRNAs respectively. EVs released by reduced mercaptopurine cytotoxicity (about 10%) and Rac1 protein expression in KD cells (ANOVA, $p < 0.001$), probably because they transport different miRNAs.
In conclusion, *PACSIN2* KD increase mercaptopurine cytotoxicity, probably, by deregulation of miRNA expression in cells and EVs. These results will be further investigated to better explain the link between PACSIN2 and EVs, whose miRNAs could provide a new scenario in personalizing thiopurine treatment.

Keywords: PACSIN2, extracellular vesicles, miRNA-sequencing, mercaptopurine

Acknowledgement: This work was supported by the Italian Ministry of Health, through the contribution given to the Institute for Maternal and Child Health IRCCS Burlo Garofolo, Trieste, Italy, grant RC 23-23.

Poster presentation

Pathway analysis of CD8+ T cell transcriptome in glioblastoma patients reveales multiple sclerosis signaling pathway as the top rated upregulated disease pathway in tumor infiltrating cells

Milan Stefanović[1]*, Ivan Jovanović[1], Aleksandra Stanković[1] and Maja Živković[1]

[1]Institute for nuclear sciences ''Vinča'', National institite of the Republic of Serbia, Laboratory for radiobiology and molecular genetics, Mike Petrovića Alasa 12-14, 11351 Vinča, Beograd, Srbija
milanst@vin.bg.ac.rs

The significance of CD8+ T cell central nervous system migration and activation in the progression of glioblastoma is well documented. However, molecular signaling pathways regulation related to migration and activation in CD8+ cells in glioblastoma is scarce. Therefore we have analyzed the molecular pathway regulation in differentially expressed mRNAs of tumor infiltrating vs. peripheral blood CD8+ T cells from glioblastoma patients. Tumor-infiltrating vs. peripheral blood differentially expressed mRNAs were obtained by analyzing the FASTAQ files on the Galaxy platform using the LimmaVoom tool with filtering low count mRNAs (CPM > 2). We used publically avaliable FASTAQ files with CD8+ T cells mRNA sequencing data deposited at NCBI's GEO database (accession number GSE171197). The differentially expressed mRNA were analyzed with Qiagen's Ingenuity pathway analysis (p adj. cutoff 0.05). Protein-protein interaction network was constructed on the NetworkAnalyst platform using the IMeX database with minimal order parameters. The top rated disease canonical pathway was the multiple sclerosis (MS) signaling pathway, with 18 differentially expressed mRNA hits (out of possible 222), p adj. = 0.0009 and Z score = 2.828, implying significant upregulation of this pathway in tumor-infiltrating CD8+ T cells.

The MS signaling pathway describes the molecular cascade which leads to the autoimmune phenotype in lymphocytes, including activation and central nervous tissue infiltration. To further specify the aspects of the canonical MS signaling pathway which might influence tumor infiltrating phenotype we have constructed a minimal order protein-protein interaction network. Results showed a number of lymphocyte migration and activation KEGG terms within the network, such as: TNF signaling pathway (p adj. = 0.0000115), IL-17 signaling pathway (p adj. = 0.00000427), sphingolipid signaling pathway (p adj. = 0.00171), NF-kappa B signaling pathway (p adj. = 0.0000694) and TCR signaling pathway (p adj. = 0.0071).

We conclude that MS signaling pathway is an viable model for further understanding of the transcriptional phenotype of glioblastoma infiltrating CD8+ T killer cells, illustrating that same migration and activation mechanisms which mediate brain autoimmunity are essential for brain antitumor adaptive immunity.

Keywords: glioblastoma, multiple sclerosis, enrichment analysis, network analysis

Transcriptomic profiling of white blood cells reveals new insights into the molecular mechanisms of thalidomide in children with inflammatory bowel disease

Marianna Lucafò[1]*, Letizia Pugnetti[2], Debora Curci[2], Carlotta Bidoli[1], Marco Gerdol[1], Fulvio Celsi[2], Sara Renzo[3], Monica Paci[3], Sara Lega[2], Paolo Lionetti[3], Alberto Pallavicini[1], Giuliana Decorti[2,4], Gabriele Stocco[2,4], Matteo Bramuzzo[2]

[1] Department of Life Sciences, University of Trieste, Trieste, 34127, Italy
[2] Institute for Maternal and Child Health, IRCCS "Burlo Garofolo", Trieste, 34137, Italy
[3] Gastroenterology and Nutrition Unit, Meyer Children's Hospital IRCSS, Florence, 50139, Italy
[4] Department of Medicine, Surgery and Health Sciences, University of Trieste, Italy
mlucafo@units.it

Thalidomide has emerged as an effective immunomodulator in the treatment of pediatric patients with inflammatory bowel disease (IBD) refractory to standard therapies. Cereblon, a component of E3 protein ligase complex that mediates ubiquitination and proteasomal degradation of target proteins, has been identified as the primary target of thalidomide. Cereblon plays a crucial role in thalidomide teratogenicity, however it is unclear whether it is also involved in the therapeutic effects in IBD patients. This study aimed at identifying the mechanisms underpinning thalidomide action in pediatric IBD. Ten IBD pediatric patients clinically responsive to thalidomide were prospectively enrolled. RNA-sequencing and functional enrichment analysis was carried out on peripheral blood mononuclear cells obtained before and after treatment with thalidomide. RNA-sequencing analysis revealed 378 differentially expressed genes after treatment with thalidomide. The most deregulated pathways were cytosolic calcium ion concentration, cAMP-mediated signaling, eicosanoid signaling and inhibition of matrix metalloproteinases. Neuronal signaling mechanisms such as CREB signaling in neurons and axonal guidance signaling also emerged. Connectivity Map analysis revealed that thalidomide gene expression changes were similar to those induced by MLN4924, an inhibitor of NEDD8 activating enzyme, suggesting that thalidomide exerts its immunomodulatory effects by acting on the ubiquitin-proteasome pathway.
In vitro experiments on cell lines confirmed the effect of thalidomide on altered candidate pathways observed in patients. These results represent a unique resource for enhanced understanding of thalidomide mechanism in patients with IBD, providing novel potential targets associated with drug response.

Keywords: RNA-sequencing, thalidomide, pediatric, Crohn's disease, ulcerative colitis

Acknowledgement: This work was supported by the Italian Ministry of Health, through the contribution given to the Institute for Maternal and Child Health IRCCS Burlo Garofolo, Trieste, Italy, grant NET-2013-02355002 and RC 10/19.

Poster presentation

The past, the present, and the future of RNA secondary structure prediction

Lazar Vasović[1]

[1]Faculty of Mathematics, University of Belgrade, Studentski trg 16,
11000 Belgrade, Serbia
pd212006@alas.matf.bg.ac.rs

RNA is a biopolymer whose primary structure is a sequence of nucleobases. While messenger RNA is probably the most known, an increasing number of non-coding RNAs is being discovered. In order to become biologically active, ncRNA folds intramolecularly, thus forming segments of paired bases. This secondary structure largely determines the function of an ncRNA, so its prediction is important for newly discovered sequences. Owing to the strong link between the two structural levels, most predictors are data-driven and sequence-based.

The oldest and simplest algorithm was base pair maximization (BPM), which did not presume important structural features. Another approach exploited the fact that biophysics dictates RNA folding, so it searched for the thermodynamically optimal structure. Statistical learning was the base of the third group, with probabilistic context-free grammars (PCFGs) being the most influential. These were the state-of-the-art methods at the beginning of the century.

However, much has changed in the last years, since technological advancement allowed the widespread use of machine learning. Its use in the RNA structure prediction ranges from being the supplementary method (e.g., for estimating thermodynamical and statistical parameters of traditional methods) to encapsulating the whole prediction process. The highest success has been reported with transformers, recurrent, and convolutional neural networks (CNN).

This paper was designed as a review and aimed to compare several methods theoretically and assess them practically. As expected, model complexity was highly correlated with accuracy. On the subset of simply structured transfer RNA, for example, BPM predicted ~22% of pairings correctly, PCFG ~86%, and CNN ~99%. Other subsets, such as 16S ribosomal RNA, were more challenging, but deep learning always performed best. With the continued growth of computational power and the amount of annotated data, prediction accuracy is expected to get even closer to the experimental determination, while still maintaining a much lower cost.

Keywords: RNA structure prediction, review, machine learning

Acknowledgement: This research was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the scholarship project for young and unemployed doctoral students, contract number 451-03-1271/2022-14/2990.

The use of tryptic food protein digests data in public proteomic repositories to assess the effects of chemical and post-translational modifications on digestion outcomes

Ivana Prodić [1]*, Teodora Đukić[2], Vesna Jovanović[3], Katarina Smiljanić[3]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Vojvode Stepe 444a, 11042 Belgrade, Serbia
[2]University of Belgrade, Faculty of Medicine, Institute of Medical Chemistry, Višegradska 26, 11000 Belgrade, Serbia
[3]University of Belgrade – Faculty of Chemistry, Center of Excellence for Molecular Food Sciences & Department of Biochemistry, Serbia;
ivana.prodic@imgge.bg.ac.rs

Porcine-derived trypsin generated proteomic data of the major peanut allergen Ara h 1 from the peanut was reassessed to search for possible facilitating/hindrance effects on trypsin digestion efficacy caused by post-translational and chemical modifications (PTMs) positioned on arginine or lysine (K/R) residues. If the potential effects caused by PTMs are observed with porcine trypsin, they can be just augmented and more pronounced within human intestinal digestion. The reasoning is in inferior performance of human trypsin compared to porcine-derived used in proteomic digestion protocols, also in the lower trypsin-to-sample ratio and much shorter digestion times, even though gastric digestion precedes and trypsin is not the sole digestive enzyme.
A novel method was developed to decipher cleavage or miscleavage outcomes at scissile bonds in each, modified and unmodified sequence counterparts, using PEAKS Studio-X+ (Bioinformatics Solutions Inc., Ontario, Canada) in the reassessment of high-resolution tandem mass spectrometry data, from 18-hour long trypsin digestion proteomic protocols. In general, eight site-specific and modified K/R residues with methylation, dihydroxy and formylation showed significantly higher content of miscleaved bonds (at least >10%) compared to their unmodified counterpart peptides. Specifically, dihydroxylation and formylation hindered trypsin efficacy, while methylation on several K/R showed opposite effects.
It is essential to elucidate the specific impacts of modifications on trypsin digestion performance and if there are additional effects generated by food processing, which could influence digestion outcomes and allergenicity of food proteins/peptides.

Keywords: Trypsin, PTMs, PEAKS, K/R residues, mass spectrometry

Funding: Ministry of Science, Innovation and Technological Development of Republic of Serbia Grants No. 451-03-47/2023-01/200168 (UBFC) and 451-03-47/2023-01/200042 (IMGGE).

Poster presentation

Machine learning-based data correlation between scanning electron microscopy images and energy-dispersive X-ray spectroscopy profiles

Ahmed Musa[1,2], Baeckkyoung Sung[1,3]*, Leon Abelmann[4]

[1] Biosensor Group, KIST Europe Forschungsgesellschaft mbH, 66123 Saarbrücken, Germany
[2] Department of Computer Science, Saarland University, 66123 Saarbrücken, Germany
[3] Division of Energy & Environment Technology, University of Science & Technology (UST), Daejeon 34113, South Korea
[4] Faculty of Electrical Engineering, Mathematics & Computer Science, Delft University of Technology, 2628 CD Delft, the Netherlands
sung@kist-europe.de

Characterisation of organic and inorganic microparticles has long been an important topic in the field of environmental health sciences. Especially, combined analytical method of scanning electron microscopy (SEM) associated with energy-dispersive X-ray spectroscopy (EDX) is a commonly exploited approach to obtain extensive data on the size, morphological, and elemental information from the particulate specimens. Particulate matter (PM) is a representative atmospheric pollutant that may exert adverse effects on the human respiratory system, and SEM-EDX is a widely used tool for extracting PM analysis data, which can be subsequently utilised as physicochemical features for toxicological predictions.

In this presentation, we show a machine learning-based automation of SEM-EDX correlation of environmental PM data. First, we segment SEM images using WEKA trainable segmentation which is based on a random forest algorithm to classify pixels as foreground and background groups, followed by finding connected components (pixels that are foreground and connected vertically or horizontally). These regions are used to calculate PM shape parameters. Next, element maps are obtained from EDX using curve fitting with HyperSpy Python package. PM regions from SEM images are utilised to sum intensities in the same spatial location for the element maps to obtain elemental profiles. We finally build two models to predict PM elements: (1) Element maps from SEM-EDX data using image-to-image translation, and (2) regression to predict PM element percentages from shape features. Results from model 1 and 2 are then applied to extract PM elemental profiles associated with PM morphology information. Our results show how to efficiently predict EDX and element maps from SEM images with a high degree of accuracy. This method has a potential to significantly reduce time and labour required for environmental PM monitoring.

Keywords: Environmental health, particulate matter, SEM, EDX, automated data analysis, multiple output regression

Protein structural differences in Cytochrome c oxidase subunit 1 of two *Heterogynis* species as a new approach for species delimitation

Marija Vidović[1*], Vladislava Galović [2]

[1]Institute of Molecular Genetics and Genetic Engineering, Laboratory for Plant Molecular Biology, University of Belgrade, Vojvode Stepe 444a, Belgrade, Serbia
[2]University of Novi Sad, Institute of Lowland Forestry and Environment (ILFE), Antona Čehova 13, 21000 Novi Sad, Serbia
mvidovic@imgge.bg.ac.rs

Insects are the most diverse group in the animal kingdom, accounting for about two-thirds of all animals. Cytochrome c oxidase subunit 1 (COI) is the most commonly used marker gene for animal species delineation. However, the accuracy of this approach crucially depends on the degree of overlap between the intra- and interspecific variations.
Recently, we have identified a new species, *Heterogynis serbica* sp. n. (Lepidoptera: Zygaenoidea, Heterogynidae) found on the Mt. Kopaonik, Republic of Serbia, Balkan Peninsula. This was done by integrating taxonomic approach using morpho-anatomical characteristics by comparative scanning electron microscopy (SEM), linear wing morphometry and COI-based DNA barcoding [1]. In this study, we have used a set of bioinformatics tools available online, to determine the differences in secondary and tertiary structure of the COI proteins from *H. serbica* sp. n. and *H. zikici*. We also compared the amino acid distribution and COI motif profiles between the two species. Our results provide strong evidence that protein structure of COI can help with COI-based DNA barcoding for taxon-specific purposes of species identification and delimitation studies. Millions of COI DNA sequences deposited in the public domain (which are still growing) carry huge potential for a comprehensive assessment of genetic variation in COI among insects by using here described analysis.

Keywords: cytochrome c oxidase subunit 1, Heterogynidae, Heterogynis sp., Lepidoptera, conserved protein motif, protein sequence, secondary and tertiary protein structure, transmembrane helices.

Acknowledgements: This work was funded by the Ministry of Science, Technological Development and Innovation, Republic of Serbia (Contract No. 451-03-47/2023-01/200042; 451-03-47/2023-01/200197).

Poster presentation

Potentially relevant variants of unknown significance in NGS-tested patients with suspected skeletal dysplasia

Marija Mijovic*[1], Goran Cuturilo[1,2], Jelena Ruml Stojanovic[1], Aleksandra Miletic[1], Brankica Bosankic[1], Hristina Petrovic[1], Bojana Vasic[1], and Nadja Vukasinovic[1]

[1]University Children's Hospital, Department of Clinical Genetics, Belgrade, Serbia
[2]Faculty of Medicine, University of Belgrade, Belgrade, Serbia
marija.mijovic1987@gmail.com

ACMG recognizes five different categories of sequence variants identified by next generation sequencing (pathogenic, likely pathogenic, variants of unknown significance, likely benign and benign). Sometimes, potentially relevant gene variants could be categorized as variants of unknown significance according to the level of available evidences. Because of that, detailed assessment of the phenotype-genotype correlation by the clinical geneticist in each individual case is crucially important. The interpretation and classification of a variant may change over time. Variant reinterpretation is defined as the practice of reevaluating all the evidence available about the pathogenicity of a genetic variant and taking into account any new evidence that is made available since the previous interpretation.

For the last seven years, we had 168 patients with clinically suspected locus heterogeneous skeletal dysplasia. Next generation sequencing (NGS) using clinical exome sequencing or whole exome sequencing was performed for all. All patients underwent detailed phenotype-genotype correlation investigation.

Molecular diagnosis by determining the pathogenic or likely pathogenic causative gene variant(s) was established for 102 out of 168 patients (60.71%). Additionally, in 10 patients (5.95%) variant of unknown significance (VUS) with good phenotype-genotype correlation was identified. These VUS variants could be potentially, and possibly are, causal, although there are no reliable evidences of their pathogenicity at the moment. In one of the positive patients in our study, the variant was initially classified as VUS, but with new evidence it was reclassified as likely pathogenic.

In the present study, a potentially relevant variant of unknown significance was detected in 5.95% of patients, which is a non-negligible proportion. For all these patients, we have organized clinical follow-up with periodic reinterpretation and reclassification of the detected variants.

Keywords: next generation sequencing, variant(s) of unknown significance, classification, reinterpretation, reclassification, skeletal dysplasia

Analysis of Long COVID Phenotypes and their Impact on Mental Health and Daily Functioning: Insights from Twitter

Marko Markovikj[1], Jovana Dobreva*[1], Mary Lucas[2], Irena Vodenska[3], Lou Chitkushev[2], Dimitar Trajanov[1,2]

[1]Faculty of Computer Science & Engineering, Ss. Cyril and Methodius University, Rudzer Boshkovikj 16, 1000 Skopje, Macedonia

[2]Computer Sciences Department, Metropolitan College, Boston University - Boston, MA, US

[3]Administrative Sciences Department, Metropolitan College, Boston University - Boston, MA, US

jovana.dobreva@finki.ukim.mk

In this study, we conducted an investigation into Long COVID from a user perspective, utilizing Twitter social media data. Prior to analysis, the data underwent preprocessing to obtain raw text per tweet. Our analysis commenced with basic statistical analysis and subsequently expanded to identify characteristic periods for the phenotypes based on dynamic timelines. We also explored the relationships between the phenotypes, as well as the interdependence between phenotypes and geolocation.

In the context of this research, an analysis was conducted on a collection of tweets that encompassed the timeframe from March 2020 to March 2022. The dataset consisted of approximately 1.9 million tweets. In order to concentrate on word phrases, extraneous elements such as mentions, emoticons, links, and hashtags were eliminated. Subsequently, a process of lemmatization was performed. For the purpose of reducing the number of distinct phenotypes under investigation and facilitating the presentation of results, the collected data was categorized into five overarching groups: Cardiovascular, Respiratory, Daily Living, Neurological and Mental Health, and Other.

The statistical data regarding the most commonly used words by individuals describing their experiences during the Long COVID period are as follows: "Ampicillin" was tweeted 125,295 times, "Death" was tweeted 121,156 times, "Suffer" was tweeted 125,113 times, and "Vaccine" was tweeted 108,968 times. We observe distinct patterns in the emergence of certain phenotypes during this period, particularly in relation to the quality of life. On August 1, 2020, the term "quality of life" was mentioned in only 223 tweets, whereas one year later, during the same month, this phenotype garnered 1,663 tweets.

Our findings reveal that the occurrence of Long COVID phenotypes is influenced by both temporal and geographical factors. The analysis shows a clear and notable trend within the dataset. Specifically, it is observed that neurological symptoms, along with symptoms that impede individuals' daily functioning, exhibit the highest prevalence, particularly during the latter half of the analyzed tweet period. This period corresponds to a time when an increasing number of individuals have recovered from COVID-19 and are reporting their experiences with Long COVID. Notably, fatigue, depression, stress, and anxiety emerge as the most prevalent phenotypes.

This scientific investigation of the complex interactions between Long COVID phenotypes, mental health, and the manifestation of diverse symptoms is offering insights into the profound consequences on individuals' lives. These findings shed light on the significant burden posed by Long COVID and its cascading effects on various aspects of individuals' well-being and society at large.

Keywords: Long COVID, data mining, computer science, nlp

Poster presentation

Metagenomic Analysis of Bacterial Community and Isolation of Representative Strains from Vranjska Banja Hot Spring, Serbia

Jovana Curcic*[1], Danka Matijasevic[1], Nemanja Stanisavljevic[1], Srdjan Tasic[2], Milan Kojic[1], and Milka Malesevic[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
[2] College of Applied Technical Sciences Nis, Nis, Serbia
jcurcic@imgge.bg.ac.rs

Thermal springs represent a habitat with extreme conditions that harbor a unique microbial community adapted to thrive in this environment. In addition to the geothermal springs in Iceland, the thermal springs of Vranjska Banja are considered the hottest in Europe with a water temperature of 63-95°C. Due to global warming and climate change, there is a growing need for knowledge about the biodiversity of extreme natural habitats. Besides the exceptional importance of studying extremophilic microorganisms, the difficulty in their cultivation limits the expanding necessity of research in this field. This study provides information about the microbial community structure and physicochemical characteristics of the thermal spring of Vranjska Banja. To determine and monitor the microbiota diversity of the Vranjska Banja hot spring, for the first time, comprehensive culture-independent metagenomic analysis in parallel with a culture-dependent approach was applied. The culture-independent composition of bacterial communities of the thermal water was investigated using MiSeq-Illumina technology and analyzed by the computing environment QIIME2 v2021. The applied cultivation approach resulted in the isolation of 17 strains belonging to genera *Bacillus, Anoxybacillus, Hydrogenophilus, and Geobacillus*, based on 16S rRNA sequencing and whole genome sequencing of five representative strains has been performed. The complete DNA was sequenced using Illumina HiSeq from the MicrobesNG service. Genomic characterization and OrthoANI analysis have shown that two of them are candidates for novel species. Products of extremophilic microorganisms adapted to harsh conditions have great potential to be used for biotechnological research and industrial application. Results of BAGEL4 and AntiSMASH showed that the sequenced strains from Vranjska Banja hot spring have the potential to produce thermostable enzymes (proteases, lipases, amylases, phytase, chitinase, and glucanase) and various antimicrobial molecules.

Keywords: 16S rRNA Metagenomic analysis, Genome analysis, Thermophilic bacterial diversity, Hot spring, Enzymatic potential, Antimicrobial molecules

ISBN: 978-86-82679-14-1