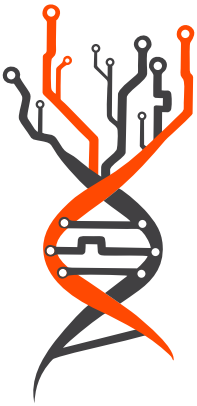
#BelBi2024 • Belgrade, Serbia

BOOK OF ABSTRACTS



5th Belgrade Bioinformatics Conference

17 - 20 JUNE **2024**

EDITORS
Dr. Ivana Morić
Dr. Valentina Đorđević

ISBN: 978-86-82679-16-5

belbi.bg.ac.rs

Title 5th 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 dr. Dušan Radojević

ISBN **978-86-82679-16-5**

Copyright © 2024 Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

BelBi2024 Committees

International Advisory Board

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

Elena Fimmel
Mannheim University of Applied Sciences, Mannheim, Germany

Guanglan Zhang,
Boston University, Boston, USA

Konstantin Severinov,
Waksman Institute for Microbiology, Rutgers University, New Jersey, USA

Lou Chitkushev,
Metropolitan College, Boston University, Boston, USA

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

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

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

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

Biljana Stanković,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia

Branislava Gemović,
VINCA Institute of Nuclear Sciences, 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

Fotis Psomopoulos,
Centre for Research and Technology Hellas, Thessaloniki, Greece

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

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

Marko Đordević,
Faculty of Biology,
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

Peter Tompa,
VIB Structural Biology Research Center,
Brussels

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

Marko Đorđević,
Faculty of Biology,
University of Belgrade

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

Jovana Kovačević,
Faculty of Mathematics,
University of Belgrade

Nenad Mitić,
Faculty of Mathematics,
University of Belgrade

Saša Malkov,
Faculty of Mathematics,
University of Belgrade

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

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

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

Tanja Vukov,
Institute for Biological Research "Siniša Stanković", University of Belgrade, Belgrade

Dušan Radojević,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

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

Nikola Kotur,
Institute of Molecular Genetics and Genetic Engineering, University of Belgrade

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

FOREWORD

We are pleased to announce the successful conclusion of the 5th Belgrade Bioinformatics Conference - BelBi2024, where numerous high-quality scientific contributions were presented. We sincerely thank all participants and proudly present a book of abstracts that not only reflects the scientific richness and diversity of the conference, but also serves as a lasting memento of this remarkable event.

This international conference was jointly organized by several research institutions, faculties, and scientific societies from Serbia. It covered a wide range of topics from the fields of computational biology, bioinformatics, biomedical informatics, and health informatics. The main goal of BelBi 2024 was to promote contacts between scientists of all levels, provide a platform for the exchange of experiences and present the latest advances in their fields. We hope that BelBi2024 was a valuable platform for researchers from all over the world to meet, build new collaborations and expand professional networks.

We are grateful and proud that we were able to welcome over 250 researchers from 21 countries from three continents. The conference included 24 scientific sessions with more than 68 oral presentations (including eight keynote lectures), 54 poster presentations, three hands-on workshops and three satellite events – the MICOS-EU competition, the TranSYS final conference and Shere the IDEA session. We also organized two industry presentations and two panel discussions - "Building Skills for the Future: Masters 4.0 in Bioinformatics" and "BIO4 Campus: Transforming Science into Business". We also presented the first BelBi art exhibition inspired by scientific discoveries, entitled "IMGGE Magnificent Cell Dance". And finally, we are particularly proud of the "Future Keynote Speakers" program, which enabled students from faculties across Serbia to attend this year's keynote lectures and panel discussions for free.

We would like to thank all the members of the International Advisory Board and the International Program Committee for their efforts and help that contributed to the success of this event. We are very grateful to the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, the SAIGE project and the Chamber of Commerce and Industry of Serbia for their support. Finally, the local organizing committee is very grateful to all sponsors of the conference - BGI & MGI, Elta'90MS, PacBio & East Diagnostics, Alfa Genetics, Vivogen, LKB, Altium, Telekom Srbija, Labena, AlphaMed, Galen Fokus, Superlab, Kefo, RNIDS, Danau Lab Beograd, RTC and Biomedica, and we hope that they will stay with us for many years to come.

Thank you once again to all who contributed to the success of BelBi2024. We look forward to seeing you at future conferences.

Warm regards,
Belgrade, July 2024

Dr. Valentina Đorđević
& Dr. Ivana Morić,
On behalf of BelBi2024
Organizing Committee

ORGANIZER

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade

MAIN CO-ORGANIZERS

Faculty of Mathematics,
University of Belgrade

Mathematical Institute of SASA,
Belgrade

Faculty of Biology,
University of Belgrade

Vinča Institute of Nuclear Sciences,
University of Belgrade

Serbian Society for Bioinformatics and Computational Biology

CO-ORGANIZERS

Institute for Biological Research "Siniša Stanković",
University of Belgrade

Institute for Medical Research,
University in Belgrade

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

Serbian Society for Microbiology

Biophysical Society of Serbia

Center for the Promotion of Science

TABLE OF CONTENT

KEYNOTE LECTURES

INVITED LECTURES

ORAL PRESENTATIONS

POSTER PRESENTATIONS

FLASH TALKS

Towards Implementing Genetic Information in Health Care and Prevention

André G. Uitterlinden

Laboratory for Population Genomics, Erasmus MC,
Rotterdam, The Netherlands
a.g.uitterlinden@erasmusmc.nl

Variance among individuals in disease susceptibility, treatment response and/or progression, is determined -in part- by genetic variation. Human genome sequencing has uncovered hundreds of millions of genetic variants, while DNA analysis technology has progressed to allow sequencing a human genome in <24hours, and to analyze millions of SNPs in millions of DNA samples using arrays.
Array technology has identified thousands of genetic factors for common disease by Genome Wide Association Studies (GWAS) in cohort studies and biobanks. Thousands of SNPs associated with disease risk for hundreds of diseases, have been combined in many disease-specific Polygenic Risk Scores (PRS). Some PRS are now evaluated in clinical trials to assess their added value in risk prediction, e.g. in population screening programs such as mammography for breast cancer. Interestingly, such PRS-es and clinically relevant DNA variants can be assessed using array genotyping (< 30 euro's/sample).
World-wide large-scale sequencing projects are now ongoing, such as the European 1 million Genomes (1MG) Project, and stimulate national genome programs. The recently funded Genome of Europe (GoE) project will target to whole genome sequence (WGS) 100,000 citizens from 29 countries across Europe. Such WGS data will be made accessible for research and applications in care and prevention, such as comparison of particular DNA variations across ethnic groups and countries, specifying the population genetic structure of subgroups in the general population, and optimizing imputation capacity against improved genomic reference data.
Together, these developments have led to genetic information now entering the hospital clinic, whereby –in theory- all patients can be assessed by array genotyping for mutations, PRS, pharmacogenetics (PGx), and blood group/HLA typing for example, to help clinicians in decision making for diagnosis and treatment, and to provide self-empowerment for patients for prevention. Such a program, called GOALL (Genotyping On ALL patients) is currently running at Erasmus MC, The Netherlands. However, also outside of the (academic) hospital setting applications of using genetic information are explored, such as in population screening programs for breast or colon cancer. Aspects of these developments and outlooks to the future will be discussed.

Keywords: genetic, sequencing, arrays, genome, PRS

Cell Types of Adult Mouse Brain: Definition and Experimental Access

Bosiljka Tasić

Allen Institute for Brain Science, Seattle, USA
bosiljkat@alleninstitute.org

The human brain is an incredibly complex organ, composed of over 150 billion cells that work together to create consciousness and ultimately, define who we are. Abnormal function or death of specialized cell types cause various brain diseases.
Understanding how brain structure produces its function is the key goal of neuroscience. To define brain structure, we need to identify the types of building blocks (cell types) and their relationships. Then we need to eliminate or inactivate them and observe the consequences (e.g., inability to perform an action, like movement). Ethical barriers prevent us from using this approach in humans.
Mus musculus, the house mouse, is a dominant model for studying mammalian brains. Despite its small size, the mouse performs diverse behaviors common across mammals, including sophisticated movements and learning.
The Allen Institute for Brain Science stands at the forefront of defining cell type identity and function in the mammalian brain. Starting with single-cell transcriptomics followed by measurements of other cellular properties including morphology and electrophysiology, we created an extensive repository of brain single-cell data. We employed various bioinformatics approaches to analyze these multidimensional and multimodal data to define cell identity and cell types. We showed that the mouse brain contains at least 5000 cell types of which many exist in humans.
To assess cell type function, we used our single-cell measurements of chromatin accessibility to define putative enhancer elements in the mouse and human genome. When included in innocuous viruses, these enhancers can instruct expression of various molecular tools in specific cell types to probe their function in the brain.
Starting with mice and with an eye towards humans, the Allen Institute is building genetic tools for all cell types in mammalian brains. Coupled with advanced computational tools, our ability to understand the roles of all brain cell types in health and disease and modify their function toward cures is within reach.

Keywords: single-cell transcriptomics, cell types, mouse brain, genomics, enhancers

Acknowledgement: This work is the result of a large team of scientists at the Allen Institute for Brain Science and its collaborators. The work was supported by the United States National Institutes of Health Grants: U19MH114830 to Hongkui Zeng, RF1MH121274 to B.T. and Tanya Daigle, RF1MH114126 and UG3MH120095 to Boaz Levi, Jonathan Ting and Ed Lein, and UF1MH128339 to B.T., Jonathan Ting, Boaz Levi, Trygve Bakken and Tanya Daigle.

Beyond 10,000 Ancient Human Genomes; Ancestral Origins at the Balkans

Carles Lalueza-Fox[1,2]

[1] Natural Sciences Museum of Barcelona
[2] Institute of Evolutionary Biology (Barcelona)
carles.lalueza.fox@gmail.com

With more than 10,000 ancient human genomes published in 2023, thanks to new technological developments on DNA sequencing, we are now able to investigate multiple ancestry layers associated to past migrations that have shaped the genomes of modern populations. These studies have been able to unravel past social structures, as well as selective processes, that left genomic marks. In the Balkans, the recent analysis of some hundreds of ancient genomes from the last three thousand years have uncovered the genetic signals of globalisation during the Roman Empire and also the signals of the Slavic migrations after the 6th century BCE. Getting into historical periods, these population movements have strong cultural and even political implications, showing the complex nature of ancestry, genetics and identity. Genetics can offer objective data on human past and yet, their interpretation in terms of identity is complex. A multidisciplinary approach, involving different disciplines such as archaeology, anthropology, history and even linguistics is recommended.

Keywords: bioinformatics, computer science, ancient DNA, ancestry, human migrations

20 years of work on exosomal DNA fragments

Christoph W. Sensen

Hungarian Center of Excellence for Molecular Medicine,
HCEMM Kft, Szeged, Hungary
christoph.sensen@hcemm.eu

In the year 1948, French authors Mandel and Metais first described the presence of cell-free DNA molecules (cfDNA) in the blood of mammals and also made a link of the quantities of this DNA fraction to various diseases. Until the advent of high-throughput DNA sequencing, this finding was mostly ignored, but since then, work on cfDNA has increased.

The Sensen laboratory has performed several studies on the DNA molecules present in plasma or serum, respectively, mostly using data obtained from paired-end high-throughput Illumina sequencing experiments, followed by Bioinformatics analysis and ultimately qPCR. This includes studies on Chronic Wasting disease in Canadian Wapitis, work on Mad Cow Disease, radiation experiments with non-lethal doses in rats as well as work on human disease conditions.

The work of more than 20 years in Canada, Austria and Hungary will be reviewed, which has ultimately led to the development of a qPCR test, which can be used to detect the onset of human sepsis up to two days before the first clinical signs. A special focus of this talk will be on the close interconnection between the medical and laboratory work and the Bioinformatics analyses, which are required for the development of a new molecular diagnostic assay, which can be performed with standard equipment already available in the clinics using non-invasive methods.

The presentation will end with an outlook on the work that is currently ongoing at the Hungarian Center for Molecular Medicine in Szeged, which is focused on the stratification of COVID-19 patients.

Keywords: bioinformatics, data mining, nucleic acids, sequencing, molecular diagnostics

Acknowledgement: The HCEMM program is funded in by EU Horizon 2020 Grant No. 739593, NRDIO Hungary National Laboratory award Project no. TKP-2021-EGA-05 and NRDIO Hungary Thematic Excellence award, Project no. 2022-2.1.1-NL-2022-00005.

Can We Rely on AI?

Desmond John Higham

University of Edinburgh, Edinburgh, UK
d.j.higham@ed.ac.uk

Over the last decade, adversarial attack algorithms have revealed instabilities in deep learning tools. These algorithms raise issues regarding safety, reliability and interpretability in artificial intelligence (AI); especially in high risk settings.
At the heart of this landscape are ideas from optimization, numerical analysis and high dimensional stochastic analysis. From a practical perspective, there has been a war of escalation between those developing attack and defence strategies. At a more theoretical level, researchers have also studied bigger picture questions concerning the existence and computability of successful attacks. We will present examples of attack algorithms in image classification and optical character recognition. We will also outline recent results on the overarching question of whether, under reasonable assumptions, it is inevitable that AI tools will be vulnerable to attack.

Keywords: stability, adversarial attack, regulation of AI

On the prediction of protein dynamics: should one be optimistic?

Frédéric Cazals

Centre Inria d'Université Côte d'Azur
frederic.cazals@inria.fr

Protein dynamics are key to protein functions, with action modes ranging from subtle motions impacting thermodynamics, to large amplitude conformational changes involved in complex multi-body mechanisms. While the prediction of (well) folded structures may be taken as an achievement in the deep learning era with Alphafold2 and the like, predicting dynamics essentially remains an open problem.
This talk will review recent work in this realm, based on novel insights on loop closure techniques coupling kinematic models in high dimensional dihedral angle spaces, and Monte Carlo Markov Chain sampling techniques of the Hit-and-Run type. Along the way, I will discuss connexions with other problems, including high dimensional volumes and densities of states, as well as mixture models in flat tori to capture couplings between torsion angles.
These ingredients will make us ponder on the opportunity to be optimistic regarding the accurate and fast prediction of protein dynamics.

Keywords: proteins, conformational changes, loop closure, sampling, Monte Carlo Markov chains.

Future directions in network biology

Tijana Milenkovic

Department of Computer Science and Engineering,
University of Notre Dame, Notre Dame, USA
tmilenko@nd.edu

Network biology, an interdisciplinary field bridging computational and biological sciences, has revolutionized understanding of cellular functions and diseases. The field, which has existed for two decades, has witnessed rapid evolution, accompanied by emerging challenges. These challenges stem from various factors, notably the growing complexity and volume of data together with the increased diversity of data types describing different scales of biological organization.

This talk will discuss some of key research areas in network biology and highlight recent breakthroughs in these areas; offer a perspective on the future directions of network biology; and touch on scientific communities, educational initiatives, and the importance of fostering diversity within the field.

Two specific research directions will be discussed. The first is on our network-of-networks analyses of multi-scale biological systems with application to protein function prediction. The function of a protein is determined by the protein's 3D structure, which also affects which other proteins the protein interacts with. Because of this, and because nodes in a protein-protein interaction (PPI) network can be represented as protein structure networks (PSNs), we modeled the integrated PPI and PSN data as a network-of-networks. We found that the multi-scale network-of-network analysis often resulted in more accurate protein function prediction than traditional single-scale analysis of PPI data alone or PSN data alone.

Second, the talk will discuss our network-based analyses of protein folding. We had proposed several approaches for modeling protein 3D structures as PSNs. Static PSNs model the whole, final 3D structure of a protein. Because the folding of a protein is a dynamic process, where some parts (3D sub-structures) of a protein fold before others, most recently, we modeled a protein as a dynamic PSN that captures these sub-structures. We evaluated our PSN models in the task of protein structural classification. We found that our PSN models outperformed state-of-the-art approaches for the same task, with dynamic PSNs being superior to static PSNs.

Keywords: network biology, protein-protein interaction networks, protein structure networks, protein function prediction, protein folding

Acknowledgement: A part of the talk will be based on our collaborative paper titled "Current and future directions in network biology" (arXiv:2309.08478 [q-bio.MN], 2023) that is co-authored by numerous experts in network biology, who initialized the discussion from the paper at the Workshop on Future Directions in Network Biology held at the University of Notre Dame during June 12-14, 2022. The workshop was supported by the U.S. National Science Foundation [grant number CCF-1941447].

Way2Drug Platform: From Biological Activity Prediction to Systems Pharmacology

Dmitry S. Druzhilovskiy, Dmitry A. Filimonov, Anastasia V. Rudik, Polina I. Savosina and Vladimir V. Poroikov*

Institute of Biomedical Chemistry, Moscow, Russia
vladimir.poroikov@ibmc.msk.ru

Global chemical space is extremely vast and finding a molecule with the required pharmacotherapeutic properties is a formidable challenge. Starting from analysis of big chemical-biological data obtained *in silico*, *in vitro*, *in vivo* and *in clinics* it is necessary to finish with one active pharmaceutical ingredient possessing the needed safety and efficacy. Combining information extracted from the curated datasets of active/inactive compounds available through World Wide Web and AI/ML tools, investigators are surfing from global to local scales in pharmaceutical R&D, enabling faster and more efficient development of new therapeutic remedies.

Way2Drug (https://www.way2drug.com/dr/) is a quickly expanding web portal focused at integrating of demanded *in silico* tools for drug discovery. Way2Drug currently hosts services for predicting the biological activities (PHARMA), toxicity (TOX), metabolism (META) and physicochemical characteristics (ADME) of drug-like molecules. All tools are freely available for non-commercial academic research.

In addition to the predictive web-services, several informational resources are available at the Way2Drug portal, including WWAD (World-Wide Approved Drugs, https://www.way2drug.com/wwad/), phytocomponents of the Russian officinal medicinal plants Phyto4Health (https://www.way2drug.com/p4h/), host gut microbiota metabolism xenobiotics database HGMMX (https://www.way2drug.com/hgmmx/).

Way2Drug is evolving as the basis for development of computational platform for efficient analysis and interpretation of the extensive biomedical and clinical data, comparative analysis of information extracted from this data to differentiate the normal and pathological states, obtaining new knowledge to identify potential pharmacological targets and biomarkers, designing potential pharmacological substances with the required properties, determining the optimal approach to therapy taking into account the individuality of patients.

Keywords: bioinformatics, chemoinformatics, data mining, computer-aided drug discovery, Way2Drug platform

Acknowledgement: The study is performed in the framework of the Program for Basic Research in the Russian Federation for a long-term period (2021-2030) (No. 122030100170-5).

Multiomic Profiling of Early Onset Colorectal Cancer

Aleksandar Krstic

School of Medicine, Systems Biology Ireland,
University College Dublin, Dublin, Ireland
aleksandar.krstic@ucd.ie

Early onset colorectal cancer (EOCRC), characterised by colorectal cancer (CRC) in individuals under 50 years of age, has witnessed a concerning global rise over the past three decades. While various risk factors such as smoking, alcohol consumption, and dietary habits have been implicated, a definitive causal link remains to be fully understood. Our research aims to explore the hypothesis that EOCRC represents a distinct clinical and molecular entity compared to Late Onset CRC (LOCRC).

Through comprehensive omics profiling encompassing proteomics, genomics, and transcriptomics, we aim to unravel the underlying mechanisms driving EOCRC. This approach not only offers insights into disease pathogenesis, but also holds promise for identifying novel biomarkers crucial for patient stratification and the development of tailored treatment strategies.

In our study, we analysed healthy and matched tumour tissues from a cohort of 80 Irish EOCRC and LOCRC patients diagnosed with microsatellite stable, resectable disease and no familial syndromes, using mass spectrometry. Pathway enrichment analysis unveiled alterations in molecular functions specific to EOCRC samples, including dysregulated Granzyme A signalling and mitochondrial dysfunction indicative of potential metabolic rewiring. Furthermore, downregulated NET signalling suggests unfavourable immunomodulation, while observed deactivation of oxidative phosphorylation, corroborated by transcriptomics analysis, may indicate a response to immune checkpoint inhibitors.

By integrating these findings with mutational profiling and clinical data, we aim to develop digital simulations capable of predicting patient outcomes and guiding personalised treatment regimens. This multifaceted approach holds promise for advancing our understanding of EOCRC and improving patient care outcomes.

Analysis of unlikely (and rare) local protein conformations

Alexandre G. de Brevern[1,*] and Nenad Mitić[2]

[1] BDSIMB Bioinformatics team, INSERM UMR_S 1134, BIGR, Université Paris Cité, Université de la Réunion, Necker Hospital, Paris, France
[2] Department of Computer Science, Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
alexandre.debrevern@univ-paris-diderot.fr

Three-dimensional (3D) protein structures underpin the biological functions that are essential to life. Access to this 3D information is of great interest for both basic and applied research. Traditionally, structures are analyzed by assigning secondary structures (helices, sheets and loops). However, this description does not allow loops to be properly described and does not provide accurate details of the fine structure of repetitive structures.
As a result, more systematic approaches to describing 3D structures have been developed, known as Structural Alphabets (SA). Within this framework, Protein Blocks (PBs) is the SA that has had the most success and application. The 16 PBs, named from PB *a* to PB *p*, are pentapeptides that can finely approximate the entire 3D structure. They have a strong sequence-structure relationship and form a grammar in which certain PBs preferentially follows a PB. There are so highly preferential transitions. Some PBs are strongly directed to two or three PBs. However, there are also rare but present transitions, i.e. present with a frequency less than 1%.
The work carried out here involved analyzing data from the Protein Data Bank to see which transitions are very common and which are rare. Secondly, the amino acid frequencies of the PBs involved in these rare transitions were compared with the frequencies classically expected to answer this simple question: Are these rare, and therefore unexpected, transitions linked to different amino acid compositions to those observed in PBs in general. Finally, a similar analysis was carried out using AlphaFold2 models of the human proteome. This work highlights the specific behavior of a number of PBs and amino acids.

Keywords: bioinformatics, data mining, computer science, protein structures, sequence – structure relationship.

Understanding the natural activation mechanism of the CRISPR-Cas immune system in *Escherichia coli*: A Computational Modeling Perspective

Anđela Rodić[1,*], Marko Tumbas[1], Jane Kondev[2],
Magdalena Đorđević[3] and Marko Đorđević[1]

[1] Faculty of Biology, University of Belgrade, Belgrade, Serbia
[2] Martin A. Fisher School of Physics, Brandeis University, Waltham, MA, USA
[3] Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia
andjela.rodic@bio.bg.ac.rs

CRISPR-Cas systems protect bacteria from viruses by using spacers, viral DNA fragments stored within the CRISPR array in the bacterial genome. These spacers are transcribed and then processed into crRNAs, guiding Cas proteins to eliminate complementary viral DNA sequences. Despite CRISPR-Cas's extensive use in biotechnology, its natural function in bacteria, especially *Escherichia coli*, is not fully understood. In *E. coli*, CRISPR-Cas activity is silenced by cooperatively bound H-NS proteins to the cas genes promoter. Viral DNA with higher AT content might sequester some H-NS, alleviating this repression. The transcriptional regulator LeuO, whose transcription is activated by BglJ-RcsB, can further activate cas genes transcription.

This study explores whether a slight reduction in H-NS levels can provide initial transcription derepression, and trigger a positive feedback loop activating the CRISPR-Cas system expression through BglJ-RcsB and LeuO, leading to rapid crRNA generation for defense against fast-replicating bacteriophages. We developed a dynamical model for crRNA expression upon infection by foreign DNA and used the Random Forest machine learning technique to identify parameters essential for achieving a sufficient crRNA increase within 30 minutes of foreign DNA entry.

A bioinformatics analysis of 16,388 viruses and their host bacteria revealed a consistent, slight increase in viral genomic AT content compared to hosts, suggesting reduced H-NS levels available for repression upon bacteriophage infection. Significant reductions in H-NS levels can rapidly induce crRNA expression, while smaller reductions require high H-NS binding cooperativity and the baseline cellular H-NS level near to the H-NS equilibrium binding constant.

Our findings indicate that CRISPR-Cas can quickly respond to fast-replicating bacteriophages. The kinetic features, which are essential for understanding the natural function of CRISPR-Cas and enhancing biotechnological applications, should inform future experimental work.

Keywords: CRISPR-Cas, Escherichia coli, H-NS, dynamic modeling, machine learning

Acknowledgement: This work is supported by the Science Fund of the Republic of Serbia (projects No. 7750294, q-bioBDS, and No. 6417603, CRISPR modelling).

Elucidating the role of cellular parabiosis in determining the preferential site of metastasis

Andrea Gelemanović*, Tinka Vidović,
Miroslav Radman and Katarina Trajković

Mediterranean Institute for Life Sciences (MedILS), Split, Croatia
andrea.gelemanovic@medils.hr

Although primary cancer can affect any part of the body, there is a striking variability in cancer prevalence across different organs. Moreover, metastases are non-randomly distributed among organs, with lung, liver, lymph nodes and bone acting as metastatic hotspots. The reasons for such inter-organ variability in the prevalence of primary cancers and metastases are insufficiently understood. We aim to understand the impact of cellular parabiosis in carcinogenesis with specific focus on its role in determining the preferential host tissue for metastasis. Our main hypothesis is that malignant phenotype can be suppressed and delayed via the paradigm of cellular parabiosis as healthy surrounding cells would complement metastatic cell and thus the „healthy homeostasis" could be maintained. To address this question we follow several research lines with aim to: 1) identify intrinsic features within a healthy organ in the absence of cancer that would predetermine such organ to become a primary cancer or metastatic host; 2) test the role of the cellular environment in the metastatic nesting, specifically, we aim to predict the preferential site of metastasis based on gene expression patterns; and 3) perform the most comprehensive and up-to-date systematic review and meta-analysis on the prevalence of metastatic sites from different primary cancers. We show that susceptibility of organs to a primary cancer or a metastasis is linked with their distinctive intrinsic features in the healthy state. In particular, while susceptibility of an organ to primary cancer is associated with the abundance of endothelial cells and atypical gene expression, susceptibility to metastases correlates with high content of immune cells and high expression of immune genes. These data shed light on some fundamental aspects of cancer biology and pave new avenues for mitigation of cancer.

Keywords: susceptibility to cancer, metastatic organotropism, cell composition, gene expression, immune cells

Toxicology Transformed: Harnessing Artificial Intelligence for Advanced Research

Bojana Stanic[1,*], Nemanja Miloševi[2], Nataša Sukur[2] and Nebojsa Andric[1]

[1] Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia
[2] Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Novi Sad, Serbia
bojana.stanic@dbe.uns.ac.rs

Over the past few decades, toxicology has made a sharp turn from an observational science focused on analyzing chemical-induced endpoints to a data-rich discipline. The amount of new data stemming from the literature, high-throughput screening (HTS) assays, omics, and other technologies is rapidly accumulating, creating a fruitful ground for the application of artificial intelligence (AI). Through machine learning (ML), deep learning, large language models, and natural language processing techniques, AI can effectively navigate this complex data landscape, deciphering patterns, elucidating toxicity mechanisms, and enhancing risk prediction.

Here, some of the applications of ML in toxicology will be presented. The ML models were used to unravel the intricate mechanisms underlying chemical-induced female infertility. The adverse outcome pathway (AOP), a theoretical concept describing biological events leading to adverse effects, was used as a backbone in developing the ML models for female infertility. Utilizing eighteen HTS bioassays, these models tracked key biological processes outlined in AOP7 – receptor binding as a molecular initiating event, and gene expression and steroid production as key events – leading to adverse outcomes. These ML models efficiently simulated and predicted perturbations in each event within toxicity pathways for novel chemicals, revealing a group of chemicals that can affect all events in the AOP, thus forming a linear molecular pathway that can lead to female reproductive disorders. The ML models were also used to assess the potency of novel chemicals in binding to the progesterone receptor and to discriminate between agonists and antagonists of this receptor.

These examples underscore the vast potential of AI in toxicology research, offering a multitude of avenues for exploration. AI has the potential to transform toxicology into a more predictive, mechanism-based, and evidence-integrated scientific discipline to better safeguard human and environmental health from chemical hazards.

Keywords: machine learning, environmental chemicals, computational toxicology

Acknowledgement: This work was supported by the Provincial Secretariat for Higher Education and Scientific Research of the Autonomous Province of Vojvodina (grant number 142-451-3533).

Higher-Order Connectivity and Synchronization Patterns in Human Connectome Networks

Bosiljka Tadić[1,2]

[1] Department of Theoretical Physics, Jozef Stefan Institute, Ljubljana, Slovenia
[2] Complexity Science Hub, Vienna, Austria
bosiljka.tadic@ijs.si

Recent advances in the physics of complex systems aim at understanding the emergence of new features at a larger scale---the central idea of physical notion of complexity, linking the features of collective dynamical behaviors with higher-order interactions. Mapping the brain imaging data onto brain networks and revealing their detailed structures enables the application of these approaches in the study of complex brain functioning. Increasing evidence shows that connections between many brain regions play a role in specific brain processes and help describe pathways of neurodegeneration. In this context, studies highlight the importance of synchronization processes and the role of central brain regions (hubs).
In this lecture, we describe some representative examples that motivated our research. We then present more details of our results obtained by mapping human connectome data onto brain networks and their higher-order structures described by simplicial complexes. Furthermore, we present a phenomenological model of the phase synchronization processes simulated on the core network consisting of simplexes of all orders around eight brain hubs. Our results reveal partial synchronization patterns explain how the brain avoids pathological states of full synchronization. Co-evolving phases at groups of nodes (brain regions) result in multifractal fluctuations of the order parameter, which measures the degree of synchrony.

Keywords: human connectome, higher-order networks, synchronization, multifractality

Acknowledgement: Supported by the Slovenian Research Agency the Program P1-0044.

From physical to biological information and the genetic code

Branko Dragovich[1,2]

[1] Institute of Physics, University of Belgrade, Belgrade, Serbia
[2] Mathematical Institute of the Serbian Academy of Sciences and Arts, Belgrade, Serbia
dragovich@ipb.ac.rs

According to the modern scientific developments, the information is getting to be a fundamental notion like space, time and matter. These four fundamental concepts are substantially interconnected and represent the basic ingredients of the universe. Being fundamental, there is no complete definition of the information. According to our intuition, we distinguish between what is and what is not information. In the present contribution, I consider information as a very special state of the matter with a definite meaning which affects the evolution of the universe as a whole or its parts. Depending on complexity of the system one can speak about physical, biological and other information.

By physical information, first of all, I mean definite values of fundamental physical constants (c –the speed of light in vacuum, h – the Planck constant, and G – the universal gravitational constant). These constants are valid in every space, at every time and for every kind of matter. Physical informtion also includes basic properties of elementary particles (mass, electric charge, spin, ...). Physical informaton should mean everything that was given at the beginning of the universe and does not change over time.

Biological information (bioinformation) is a very special state of the biological system, which was given at the beginning of life or became during evolution. A basic example of a biological information system is DNA, which is a special long sequence of pairs of nucleotides. A part of DNA codes proteins, while the other one should be related to the regulation functions. The DNA contains special sequences of codons to which certain sequences of amino acids correspond. The special connection between 64 codons (elements of mRNA) and 20 amino acids (building blocks of proteins) with the stop signal is known as the genetic code.

In this talk, I will speak about physical and biological information as well as about some modeling of the genetic code.

"Pan-viral" disease mechanisms unveil the sweet spot for therapeutic intervention

Carme Zambrana[1,*], Sam Windels[1], Noël Malod-Dognin[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. Lluis Companys 23, 08010 Barcelona, Spain
carme.zambrana@bsc.es

Viral infections continue to cause pandemics. Antiviral drugs are based on two strategies: targeting the viral proteins or the host proteins. The main limitations of antiviral drugs targeting viral proteins are their high ratio of drug resistance and their specificity to one virus. A rapid and effective way to overcome these limitations is by re-purposing existing drugs that target the host biological mechanisms used by the virus. Moreover, when these biological mechanisms are shared across different viruses (i.e., "pan-viral" disease mechanisms), the re-purposed drugs are good candidates for broad-spectrum viral treatments.
In our previous work, we investigated how the viral interactors (VIs) of SARS-CoV-2 and differentially expressed genes (DEGs) after COVID-19 infection are connected in the host interactome. We uncover that, in the host interactome, VIs and DEGs, while mostly disjoint, are indirectly connected by their neighbours (we termed them "common neighbours" (CN) genes). Furthermore, we found that the CN genes are key to COVID-19 mechanisms and promising targets for drug re-purposing.
Here, we expand our results to 13 viruses, uncovering "pan-viral" genes and re-purposing drugs for broad-spectrum viral treatments. First, we identify CN genes for five well-studied viruses and approximate their identification for viral infections without DEG data, uncovering disease genes for 13 viruses (8 without DEG data). Then, we find that the CNs are shared across viruses with significant enrichment in viral and immune system-related processes, showing our methodology's capability to uncover "pan-viral" disease mechanisms. Finally, we predict new drug-target interactions to identify drugs targeting our "pan-viral" genes by using a two-step machine-learning model. The CNs allowed us to re-purpose drugs for treating viral infections by disrupting "pan-viral" disease mechanisms, paving the way for broad-spectrum drug re-purposing. Moreover, the CN genes can enable insight into other infectious diseases.

Keywords: omics data fusion, drug re-purposing, network biology

Acknowledgement: This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, Grant Agreement No 860895, the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grant PID2022-141920NB-I00 / AEI /10.13039/501100011033/ FEDER, UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR 01536.

Antimicrobial Rhenium Tricarbonyl Complexes: Accelerating their Discovery by Leveraging Machine Learning Models

Miroslava Nedyalkova[1], Gozde Demirci[1], Youri Cortat[1], Kevin Schindler[1], Fatlinda Rhamani[1], Justine Horner[2], Aurelien Crochet[1], Aleksandar Pavić[3], Olimpia Mamula Steiner[2], Fabio Zobi[1,*] and Marco Lattuada[1]

[1] University of Fribourg, Fribourg, Switzerland
[2] University of Applied Sciences Western Switzerland HES- SO, Fribourg, Switzerland
[3] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
fabio.zobi@unifr.ch

The expanded prevalence of resistant bacteria and the inherent challenges of complicated infections highlight the urgent need to develop alternative antibiotic options. Through conventional screening approaches, the discovery of new antibiotics has proven to be challenging. Anti-infective drugs, including antibacterials, antivirals, antifungals, and antiparasitics, have become less effective due to the spread of drug resistance. In this work, we helped define the design of next-generation antibiotic analogs based on metal complexes. For this purpose, we used artificial intelligence (AI) methods, demonstrating superior ability to tackle resistance in Gram-positive and Gram-negative bacteria, including multidrug-resistant strains. The existing AI approaches' bottleneck relies on the current antibiotics' structural similarities. Herein, we developed a machine learning approach that predicts the minimum inhibitory concentration (MIC) of Re-complexes towards two S. aureus strains (ATCC 43300 - MRSA and ATCC 25923 - MSSA). A Multi-layer Perceptron (MLP) was tailored with the structural features of the Re-complexes to develop the prediction model. Although our approach is demonstrated with a specific example of rhenium carbonyl complexes, the predictive model can be readily adjusted to other candidate metal complexes. The work shows the application of the developed approach in the de novo design of a metal-based antibiotic with targeted activity against a challenging pathogen.

Keywords: antibiotic, rhenium complexes, machine learning

Acknowledgement: M.L. and M.N. acknowledge financial support from the Swiss National Science Foundation through the NCCR Bio-inspired materials. M.L., F.Z., G.D., Y.C., A.C. acknowledge financial support from the University of Fribourg. K.S. and F.Z. acknowledge financial support from the Swiss National Science Foundation grant number 200021_196967. O.M.S. and J.H. acknowledge financial support from the Haute Ecole Spécialisée de Suisse Occidentale.

FAIR Research Software as the catalyst for trustworthy AI in Life Sciences

Fotis Psomopoulos

Institute of Applied Biosciences (INAB), Centre for Research and Technology Hellas (CERTH), Thessaloniki, Greece
fpsom@certh.gr

As a result of many years of global efforts, ensuring that data are Findable, Accessible, Interoperable and Reusable (also known as FAIR) is nowadays a clear expectation across all science domains. While data and data management have been the primary focus across many activities, research software has only recently started getting similar attention. As a result, a coordinated effort by the wider community allowed to redefine and extend the FAIR principles to research software, with similar activities now in progress aiming to enhance reproducibility, quality assurance, and long-term sustainability in software development.

At the same time, we see the emergence of the field of artificial intelligence (AI) and machine learning (ML) as a key technology impacting all sciences. As AI algorithms and models become increasingly integrated into scientific workflows, there is an urgent need to maintain high standards for research software, with the reliability and quality of the underlying software being of primary concern.

High-quality research software is definitely a key catalyst in that direction. In this context, "quality" involves not only creating robust and efficient algorithms, but also implementing rigorous quality control processes throughout the software lifecycle. There are several initiatives (such as ReSA, Turing Way and SSI) that are making available best practices, guidelines and recommendations on research software, from design and coding to testing and deployment, as well as major funded projects (such as EVERSE).

Another key aspect is around benchmarking, as it serves as a critical tool for evaluating performance, scalability, and generalizability of AI solutions across diverse datasets and use cases. In order to effectively run a benchmarking process, it is essential to establish standardized benchmarks and evaluations protocols, as well as the respective underlying services and infrastructure to facilitate this. In both cases, input and direct involvement of the respective community is essential, in order to fostering transparency and comparability in AI research.

Finally, beyond the technical aspects, there is a clear need for a coherent effort towards the interpretation of the actual FAIR principles for ML. Some efforts already exist, such as the RDA FAIR4ML interest group, as well as the efforts under the AI4EOSC project and the ELIXIR infrastructure. However, we still have some way to go, and direct community involvement is critical to ensure both wide adoption and ultimately uptake of these practices.

Keywords: bioinformatics, research software, machine learning, FAIR principles, software quality

Acknowledgement: This work has been written with the support of the EVERSE project, funded by the Horizon Europe Framework Programme (HORIZON-INFRA-2023-EOSC-01- 02) under grant agreement number 101129744. The views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. This work was also supported by ELIXIR, the research infrastructure for life science data.

Searching for the baseline miCRObiota

Antonio Starčević, Janko Diminić and Jurica Žučko*

University of Zagreb Faculty of Food Technology and Biotechnology,
Zagreb, Croatia
jzucko@pbf.unizg.hr

Microbiota, a complement of microorganisms living on and in us, has recently been correlated with plethora of health outcomes and health in general. To fully understand microorganisms with detrimental effects on human health it is necessary to establish a healthy "baseline" of the population. In our previous studies investigating correlation of obesity and gut microbiota composition and maintenance of microbiota equilibrium *in vitro* it became evident that publicly available microbiota data is lacking, especially its metadata, and it would be beneficial to create baseline microbiota composition of local population with metadata of interest. For that purpose, a pilot study of healthy working-age Croatian population was carried out. In the pilot study 60 volunteers aged between 18 and 65 years had participated providing faecal samples, for determining gut microbiota composition, and filling an online questionnaire about relevant health parameters, lifestyle, dietary habits and adherence to the Mediterranean diet. The composition of participant's microbiota did not provide novel insights, as composition on the higher taxonomic levels was similar to existing data, but it provided on finer resolution and its linkage with collected metadata. Implementation of the project also gave important feedback about weaknesses of the study and the challenges which will be addressed in the follow-up research. Results of the study are planned to be freely available, together with all anonymized metadata.

Keywords: gut microbiota, 16S rRNA sequencing, biomarkers, metadata

The genetic pathways of ferroptosis-related processes in multiple sclerosis

Maja Živković

Institute of Nuclear Sciences ''Vinča'', National Institite of
the Republic of Serbia, University of Belgrade
majaz@vin.bg.ac.rs

Multiple sclerosis (MS), a chronic inflammatory and neurodegenerative disease with no current cure, in its aetiology comprehend: susceptibility of central nervous system (CNS) to oxidative damage, mitochondrial dysfunction, impaired iron metabolism, which all lead to ferroptosis.
Therapeutic capacities to modulate ferroptosis, recently discovered cell death, have been highlighted, *in vitro*, and require further both bionformatic and experimental research to complement lack of studies in human neurodegenerative diseases. By investigating entire transcriptome in MS patients, we have identified enrichment of the Ferroptosis pathway in DEGs, before clear experimental evidence of its role in MS were presented. Consequently, the FerroReg project aimed to further investigate transcriptional and post-transcriptional regulation of ferroptosis related processes in MS. There is still a lack of specific molecular/ genetic markers that reflect ferroptosis-related molecular changes. A curated assemblage of the ferroptosis-related genes has been performed to create custom panel of 138 genes for targeted mRNA sequencing. The genes were classified according to their roles in relevant processes: lipid oxidative metabolism, antioxidant defence and iron metabolism, next to their proposed direct/indirect effect on ferroptosis. Additionally, 14 encoded transcription regulators which were associated with ferroptosis were included. Ferroptosis draws attention as a biological pathway that could be modulated, to achieve reduction of both inflammation and neurodegeneration in the central nervous system. Accordingly, gene expression was analysed with regard to disease severity, taking into account the disease modifying therapy. Applied approach overcomes the limitation of previous bioinformatic studies, which lacked the clinical data in existing gene expression data sets. Among identified DEGs, 18 genes were upregulated while 8 genes were downregulated in progressive patients compared to mild phenotype. The enrichment analysis performed on the minimum network confirmed the strong enrichment of the ferroptosis pathway while two DEGs were classified as hub molecules: TP53 and CDKN1A.

Keywords: ferroptosis, multiple sclerosis, targeted RNA-seq, custom RNA-seq panel, network analysis

Acknowledgement: This research was funded by the Science Fund of the Republic of Serbia, grant number #Grant no. 7753406, identification and functional characterization of extracellular and intracellular genetic regulators of ferroptosis related processes in multiple sclerosis, FerroReg and the Serbian Ministry of Science, Technological Development, and Innovation, Grant No. 451-03-66/2024-03/ 200017

Evidence of widespread hemizygosity and gene presence/absence variation in invertebrate pangenomes: are we overlooking the impact of genomic structural variation in metazoans?

Marco Gerdol[1,*], Nicolò Fogal[2], Carmen Federica Tucci[2], Samuele Greco[1], Marco Sollitto[3], Amaro Saco[4], Daniela Eugenia Nerelli[1], Dona Kireta[1], Magalì Rey-Campos[4], Rui Faria[5], Beatriz Novoa[4], Antonio Figueras[4], Umberto Rosani[2] and Alberto Pallavicini[1]

[1] Department of Life Sciences, University of Trieste, Trieste, Italy
[2] Department of Biology, University of Padova, Padova, Italy
[3] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Koper, Slovenia
[4] Instituto de Investigaciones Marinas (IIM), Consejo Superior de Investigaciones Científicas (CSIC), Vigo, Spain
[5] CIBIO, University of Porto, Porto, Portugal
mgerdol@units.it

Recent advancements in genome sequencing technologies have unveiled unprecedented insights into the genomic makeup of non-model metazoans, whose distinct traits have been long overshadowed by the emphasis placed on monoploid reference assemblies. While the phenotypic effects of structural variation (SV) are well-documented in plants, their significance in the animal kingdom has only recently come to light. Bivalve mollusks, with their complex genomes, present an intriguing model for investigating these phenomena. Our research on the Mediterranean mussel and on the Pacific oyster uncovered an unexpected pangenomic organization, characterized by a plethora of dispensable genes exhibiting Presence/Absence Variation (PAV) and linked to hemizygous regions. Notably, these genes were disproportionately associated with immune response and survival functions, hinting at their role in local adaptation. Subsequent examinations revealed that widespread hemizygosity is a common feature among various bivalve species, indicating an underappreciated functional impact of SV and PAV in these organisms. Interestingly, preliminary analyses suggest that gene PAV is prevalent to varying degrees in aquatic environments, underscoring the necessity for a shift in genomic research focus in non-model metazoans, from monoploid reference assemblies to pan-genomes.

Keywords: presence/absence variation; local adaptation; pangenome assembly; whole genome resequencing

Multi-scale modeling uncovers 7q11.23 copy number variation-dependent alterations in ribosomal biogenesis, mTOR and neuronal maturation and excitability

Marija Mihailovich[1,2,3,*], Pierre-Luc Germain[1,4], Reinald Shyti[1,2], Davide Pozzi[5,6], Roberta Noberini[1], Yansheng Liu[7], Davide Aprile[2,8], Flavia Troglio[1,2,8], Tiziana Bonaldi[1,8], Rudolf Aebersold[7], Michela Matteoli[5,6] and Giuseppe Testa[1,2,8]

[1] IEO, European Institute of Oncology IRCCS, Milan, Italy
[2] Human Technopole, Milan, Italy
[4] Computational Neurogenomics, D-HEST Institute for Neuroscience, ETH Zürich, Switzerland
[5] Department of Biomedical Sciences, Humanitas University, Milan, Italy
6 IRCCS Humanitas Research Hospital, Milan, Italy
[7] Department of Biology, Institute of Molecular Systems Biology, Federal Institute of Technology (ETH) Zurich, Zurich, Switzerland
[8] Department of Oncology and Hemato-Oncology, University of Milan, Milan, Italy
[3] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
marija.mihailovic@imgge.bg.ac.rs

Copy number variation (CNV) at 7q11.23 causes Williams-Beuren (WBS) and 7q microduplication syndrome (7Dup), neurodevelopmental disorders featuring intellectual disability accompanied by symmetrically opposite neurocognitive features. Although significant progress has been made in understanding the molecular mechanisms underlying 7q11.23-related pathophysiology, the propagation of CNV dosage across gene expression layers and their interplay remains elusive. Here, we uncovered 7q11.23 dosage-dependent symmetrically opposite dynamics in neuronal differentiation and intrinsic excitability. By integrating transcriptomics, translatomics and proteomics of patient-derived and isogenic induced neurons, we found that genes related to neuronal transmission follow 7q11.23 dosage and are transcriptionally controlled, while translational factors and ribosomal genes are post-transcriptionally buffered. Consistently, we found phospho-RPS6 (pRPS6) down-regulated in WBS and up-regulated in 7Dup. Surprisingly, phospho-4EBP (p4EBP) was altered in the opposite direction reflecting dosage-specific changes in the total 4EBP levels. This highlights both different dosage-sensitive deregulations of the mTOR pathway as well as distinct roles of pRPS6 and p4EBP during neurogenesis. Our work demonstrates the importance of multi-scale disease modeling across molecular and functional layers and uncovers the pathophysiological relevance of ribosomal biogenesis in a paradigm pair of complex neurodevelopmental disorders and uncouples the roles of pRPS6 and p4EBPs as mechanistically actionable relays in neurodevelopmental disorders.

Spectral Clustering for Transcriptomics Data: an Approximate Column Sampling Approach on the GPU

Marko Mišić*, Lazar Smiljković and Predrag Obradović

University of Belgrade, School of Electrical Engineering, Belgrade, Serbia
marko.misic@etf.bg.ac.rs

Clustering is one of the central methods in bioinformatics processing pipelines. It is used to capture hidden patterns in high-dimensional data, especially at the genomic level where large quantities of gene expression data have been produced with the advent of high-throughput sequencing technologies. Clustering reveals natural structure in the data, helping in understanding gene function, regulations, and cellular processes, cell assignment and subtyping, and other downstream analyses.
Spectral clustering has been successfully used to discover structure and patterns in data for bioinformatics research in the past with short execution times and high accuracy. However, it can be computationally expensive for large-scale datasets that are present in contemporary single-cell RNA (sc-RNA) and spatial transcriptomics (ST) applications. Numerous approximate spectral clustering algorithms have been proposed in the open literature to improve efficiency and scalability, while maintaining clustering quality.
Spectral clustering uses linear algebra operations which can be efficiently implemented on modern central processing units (CPUs) and graphics processing units (GPUs). In our work, we implemented an approximate, parallel spectral clustering method based on column sampling and the Nystrom method on the GPU. Our implementation significantly reduces both the computational and memory requirements of the previous methods, enabling the method to handle datasets of more than 10^6 samples and thousands of features.
We evaluated our approach using general datasets containing images of handwritten digits, as well as several datasets from single-cell and spatial transcriptomics sequencing of mouse brain. Datasets from sc-RNA and ST domains were characterized by a modestly large number of points and a relatively higher number of clusters compared to the general datasets. We compared our solution to the widely used Leiden algorithm in terms of speedup, ARI and NMI score. Our results showed significant time advantage and scalability over Leiden algorithm of up to a hundred times, albeit with to some extent lower ARI and NMI scores.

Keywords: bioinformatics, approximate spectral clustering, GPU programming, transcriptomics

Acknowledgement: This work was financially supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia under contract numbers: 451-03-65/2024-03/200103 and 451-03-66/2024-03/200103, and Complete Genomics, part of MGI Tech, under contract number: 1847/2022-12. The authors gratefully acknowledge the financial support.

Just feed it to the network - what can go wrong? Towards AI-supported early cancer detection from cytological image data

Nataša Sladoje

Centre for Image Analysis, Department of Information Technology, Uppsala University, Uppsala, Sweden
natasa.sladoje@it.uu.se

Cost-effective procedures for sample collection, as well as for their subsequent analysis, can enable large scale screening programs for early detection of oral cancer, leading to significantly improved patient survival. Brush samples taken from patients' oral cavities provide a solution to the first requirement (cost-effective sample collection). AI-supported analysis of the whole slide images of Papanicolaou stained cytological samples has the potential to offer a solution to the second (cost-effective analysis).
Can we simply train a deep neural network to classify the acquired whole slide images into two classes, differentiating samples from patients with oral cancer and samples from healthy patients?
It appears that there are many challenges to address along that road. In this talk, I will share our experience in developing methods for an AI system that can provide reliable support for a cytologist and enable interpretable, fast and cost-effective early detection of oral cancer in digital pathology.

Keywords: Digital pathology, Deep learning, Image Analysis, Whole Slide Imaging, Cancer detection

Acknowledgement: The presented work is a collaboration between a number of researchers and institutions. Acknowledgements go to J. Lindblad, J.-M. Hirsch, N. Koriakina, J. Öfverstedt, W. Lian, S. Chatterjee, C. Runow Stark, K. Edman, all from Uppsala University, and V. Bašić from Jönköping University.
The Swedish Research Council (grants 2022-03580 and 2017-04385), Sweden's Innovation Agency (VINNOVA) (grants 2017-02447, 2021-01420, and 2020-03611), and Cancerfonden (grants 22 2353 Pj and 22 2357 Pj) are acknowledged for their financial support.

The axes of biology: a novel axes-based network embedding paradigm to decipher the functional mechanisms of the cell

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

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

Common approaches for deciphering biological networks involve network embedding algorithms. These approaches strictly focus on clustering the genes' embedding vectors and interpreting such clusters to reveal the hidden information of the networks. However, the difficulty to unambiguously cluster the embeddings of genes in space and the limitations of the functional annotations' resources hinder the identification of the currently unknown cell's functioning mechanisms from the gene clusters.
We propose a new approach that shifts this functional exploration from the embedding vectors of genes in space to the axes of the space itself. We assign interpretable and fine-grained semantic meanings to the axes (basis vectors) that span the embedding space to identify the functional mechanisms of a cell.
Our axes-based methodology captures 1.32 times more functional information (GO BP terms associated with the axes) from the embedding spaces than the standard gene-centric approach (GO BP terms enriched in at least one gene cluster). This captured information is also better stratified, as GO BP terms associated with the same axis are, on average, 1.42 times more semantically coherent than those enriched in the same gene cluster. Moreover, it uncovers new data-driven functional interactions that are unregistered in the functional ontologies, but biologically coherent. We exploit these interactions to define new higher-level annotations that we validate through literature curation. Finally, we leverage our methodology to discover evolutionary connections between cellular functions and the evolution of species.

Keywords: network biology, network embedding, AI

Acknowledgement: This work is supported by the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grant PID2022-141920NBI00 / AEI /10.13039/501100011033/ FEDER, UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR01536.

What do amyloidosis, antimicrobial peptides, and the Spike RBD of SARS-CoV-2 have in common?

Oxana Galzitskaya[1,2]

[1] Gamaleya Research Centre of Epidemiology and Microbiology, 123098 Moscow, Russia
[2] Institute of Protein Research, Russian Academy of Sciences, 142290 Pushchino, Russia
ogalzit@vega.protres.ru

Some proteins and peptides are known to have antimicrobial and amyloidogenic properties. Understanding the mechanism of aggregation can be used to combat bacterial and viral diseases.
Using modern data on the process of formation of amyloid structures, we have developed and successfully tested amyloidogenic antimicrobial peptides (AAMPs) with a new mechanism of antimicrobial action - "protein knockout". This mechanism is based on the principle of directed coaggregation of AAMP and bacterial ribosomal protein S1. The innovative peptide interacts with the target protein of model or pathogenic bacteria, forming aggregates and removing this protein, which is essential for the life of the bacteria, from its working state. During the work, the antimicrobial effects of the developed peptides were examined on two model organisms (*Thermus thermophilus* and *Escherichia coli*) and two pathogenic microorganisms (*Staphylococcus aureus* and *Pseudomonas aeruginosa*).
Since the emergence of the original variant in Wuhan in 2019, many different variants of SARS-CoV-2 have been described and characterized, varying in transmissibility and pathogenicity in the human population, although the molecular basis of this difference remains controversial. Thus, the Omicron variant is known for its contagiousness, transmissibility, and lower pathogenicity (mortality). A significant role in this is played by amino acid substitutions on the surface of the Spike protein, which interact with the ACE2 receptor, which can facilitate the penetration of the virus into the cell or help it to evade the immune response. Mutations in this strain result in increased amyloidogenicity of the Omicron strain in the ACE2 receptor binding regions, resulting in an increase in the strength of this interaction for the Omicron BA.1 RBD compared to the Wuhan-Hu-1 or Delta RBD, and this effect was more pronounced at pH 5. This result is associated with Omicron variants' increased ability to spread through the population.

Keywords: amyloids, antimicrobial peptide, amyloidogenic regions, spike protein

Graphlet-based higher-order network embeddings: the past, the present and the future

Sam F. L. Windels[1], Noël Malod-Dognin[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
natasha@bsc.es

At a high level, there exist two approaches for mining networks. Neighbourhood-based approaches uncover groups (i.e., clusters) of tightly connected neighbouring nodes in a network and make predictions based on guilt by association: two nodes are assumed to be more likely to interact or share attributes if they belong to the same group(s) in the network. Topology (i.e., structure) based approaches make predictions based on structural similarity. The state-of-the-art methods to quantify local topology are based on *graphlets*, which are small connected non-isomorphic induced subgraphs. To combine neighbourhood and graphlet-based information, we defined graphlet adjacency, which weighs the adjacency of two nodes based on their co-occurrence frequency on a given graphlet (i.e., there is one type of adjacency for each graphlet). In this talk, we provide an overview of various methodologies we generalised using graphlet adjacency, including graphlet spectral embedding, graphlet eigencentrality, graphlet diffusion and hyperbolic graphlet coalescent embedding, and show how we applied them to better describe the functional organisation of various molecular networks and to better capture disease mechanisms. Recently, we used graphlet based symmetries to improve random walk based approaches. We conclude by presenting future research directions for new graphlet adjacency-based methods and applications.

Keywords: higher-order network topology, network biology, data mining

Acknowledgement: This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, Grant Agreement No 860895, the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grant PID2022-141920NB- I00 / AEI /10.13039/501100011033/ FEDER, UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR 01536.

Feature Selection For Multi-Source SCT Data

Saša Malkov* and Nenad Mitić

Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
sasa.malkov@matf.bg.ac.rs

Bioinformatics experiments often produce large data sets, with lots of samples and many different attributes, which provide high dimensionality of the data. Even if some of the dimensions have little significance for specific data analysis, they can prove useful in complex data processing. Advanced data mining techniques and AI algorithms typically welcome high dimensionality of data. However, if there are too many dimensions, we can run into the *curse of dimensionality*, because an abundance of dimensions can introduce additional complexity and cost to data handling and processing, as well as overfitting the model to less important dimensions. In order to make data processing more efficient and improve the quality of created models, we often need dimensionality reduction.

Single cell transcriptomics (SCT) is one of the most important sequencing technologies today. It enables simultaneous measurement of the activity of thousand of genes in individual cells, resulting in RNA profiles of the cells. Such profiles allow researchers to analyze the physiological activity of cells in different circumstances, including different biochemical conditions but also different health conditions of the subjects. By processing large numbers of cells, SCT provides a lot of sample data. Each detectable RNA represents one dimension of data, which ultimately gives us thousands of dimensions. Moreover, initial cells conditions, complex cell preparation techniques, and RNA measurement methods can vary significantly, resulting in significant differences in data coming from different sources.

Here we discuss the feature selection problem on the example of more than 120,000 instances of peripheral blood mononuclear cell (PBMC) SCT data from four different sources, with the detection of 30,698 genes (dimensions). We pay special attention to the imbalanced nature of the data and consider feature selection methods that allow for an unbiased set of significant features to be obtained as a result. We show that statistical correlation-based feature selection, with some support from mutual information-based techniques, can result in a reasonably complex method for high-quality feature set selection.

Keywords: bioinformatics, feature selection, statistical correlation, mutual information, transcriptomics data.

Modeling of the Hypothalamic-Pituitary-Adrenal Axis dynamics by Stoichiometric Networks

Stevan Maćešić[1,*], Ana Ivanović-Šašić[2] and Željko Čupić[2]

[1] University of Belgrade, Faculty of Physical Chemistry, Belgrade, Serbia
[2] University of Belgrade, Institute of Chemistry, Technology and Metallurgy, Department of Catalysis and Chemical Engineering, Belgrade, Serbia
stevan.macesic@ffh.bg.ac.rs

The hypothalamic-pituitary-adrenal (HPA) axis is a neuroendocrine system that regulates the body's response to stress and maintains homeostasis through the secretion of cortisol, its primary hormone. Dysregulation of the HPA axis is implicated in numerous stress-related disorders, including obesity, depression, chronic pain, metabolic disorders, etc. Therefore, understanding the HPA axis is vital for comprehending stress-related diseases and developing effective interventions. Investigating the dynamic nature of HPA axis activity presents significant challenge, which can be effectively addressed through mathematical modelling. Modelling can provide deep insights into the system's responses to stress, regulatory mechanisms involving ultradian and circadian rhythms, feedback loops, and hormonal interactions. Furthermore, modelling the HPA axis facilitates understanding how various factors influence its functioning, offering a powerful tool for studying related disorders and developing targeted interventions. Hence, this paper presents a detailed mathematical modelling approach utilizing stoichiometric networks to describe the dynamics within the HPA axis. The model captures the interplay of response strategies in the HPA axis, providing a framework for simulating its behaviour under different conditions. This model has potential for studying stress modulation, improving stress management strategies, and addressing health outcomes related to HPA axis dysregulation.

Keywords: hypothalamic-Pituitary-Adrenal Axis, HPA, stoichiometric networks, biological networks

Acknowledgement: We are grateful to the financial support from Ministry of Science, Technological Development and Innovation of Republic of Serbia (Contract numbers 451–03–66/2024–03/200026 and 451–03–66/2024–03/200146). This research was supported by Science Fund of Republic of Serbia #Grant Number. 7743504, NES. We are also especially grateful to professor Ljiljana Kolar-Anić (1947–2023), who initiated this research.

Data analysis and modelling of climate and environmental drivers of vector borne diseases - some methodological approaches and challenges of OneHealth data

Suzana Blesić

Institute for Medical Research, University of Belgrade, Belgrade, Serbia
blesic.suzana@gmail.com

Due to climate change and environmental degradation the spread and the connected risk of vector-borne diseases are spatially shifting. This requires better understanding and more detailed modelling of climate and environmental drivers of those diseases, to assist efficient and timely preparedness of primarily public and veterinary health systems for such changes, true applications for dedicated information like early warning systems. Here several modelling approaches and data analysis techniques that can be used for such purposes will be presented. These are employed in two projects – CLIMOS and PLANET4HEALTH. Finally, experiences and challenges of use of One Health datasets of sand fly borne diseases and mosquito borne diseases will be shortly discussed.
The CLIMOS consortium is co-funded by the European Commission grant 101057690 and UKRI grants 10038150 and 10039289. The six Horizon Europe projects, BlueAdapt, CATALYSE, CLIMOS, HIGH Horizons, IDAlert, and TRIGGER, form the Climate Change and Health Cluster.
The PLANET4HEALTH project is funded by European Commission grant 101136652. The five Horizon Europe projects, GO GREEN NEXT, MOSAIC, PLANET4HEALTH, SPRINGS, and TULIP, form the Planetary Health Cluster.

Keywords: vector borne diseases, data analysis and modelling, climate change and environmental degradation, One Health

Challenges in metagenomic annotation of antibiotic resistome

Svetlana Ugarcina Perovic[1,2], Vedanth Ramji[3,4], Hui Chong[1], Yiqian Duan[1], Rémi Gschwind[5], Etienne Ruppe[5,6], Finlay Maguire[7] and Luis Pedro Coelho[1,4]

[1] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China
[2] Department of Cellular, Computational and Integrative Biology - CIBIO, University of Trento, Trento, Italy
[3] APL Global School, Chennai, India
[4] Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology (QUT), Translational Research Institute, Woolloongabba, Queensland, Australia
[5] INSERM, University of Paris, IAME - Paris, France
[6] AP-HP, Bichat-Claude Bernard Hospital, Bacteriology Laboratory, Paris, France
[7] Faculty of Computer Science, Dalhousie University, Halifax, Canada
svetlana.ugarcina@gmail.com

The antibiotic resistance genes (ARGs) in both host-associated and environmental microbiomes – antibiotic resistome – play an important role in the spread of antibiotic resistance. Metagenomics enables high-throughput exploration of microbiomes. For ARGs annotation within the metagenomes, several tools are in use, however, most of them are developed for genomics studies and their databases may pose certain biases.
We compared outputs from different ARGs annotation approaches: we ran >13 000 high-quality metagenomes from 14 habitats (Coelho et al., 2022) in three input modes (read/assembly/protein) through six ARGs annotation pipelines using RGI, ABRicate, ResFinder, AMRFinderPlus, DeepARG and ARGs-OAP. To facilitate comparison of outputs with different gene names, we performed ARO (CARD's Antibiotic Resistance Ontology) normalization using our tool https://github.com/BigDataBiology/argNorm.
DeepARG and RGI provide higher coverage (with potential novel ARGs and/or false positives) while ABRicate's different databases have lower coverage but more well validated ARGs. The annotations did not differ only in hits number but also in information provided: e.g. using ABRicate, more sulfonamide ARGs were identified using ResFinderFG version 2.0 as the database (Gschwind et al., 2023), while using ResFinder resulted in more macrolide ARGs. Thus, choice of annotation tool and database should be driven by research questions and ARGs targets.
Coelho, L.P., Alves, R., del Río, Á.R. et al. Nature (2022). doi: 10.1038/s41586-021-04233-4
Gschwind, R., Ugarcina Perovic, S., Weiss, M. et al. Nucleic Acids Research (2023). doi: https://doi.org/10.1093/nar/gkad384

Keywords: microbiome, metagenome, antibiotic resistance gene

Acknowledgement: This work was supported by the International Development Research Centre, IDRC, (under the framework of JPI AMR, grant 109304-001, EMBARK) and the Shanghai Municipal Science and Technology Major Project (grant 2018SHZDZX01) (SUP, LPC); the Australian Research Council (grant FT230100724) and National Health and Medical Research Council (under the framework of JPI AMR, grant 2031902, SEARCHER) (LPC); and Emergent Ventures 2023 Fellowship (VR).

Alternatively spliced exons manifest coordinated multi-domain alteration in synapse specific genes

Vladimir Babenko

Institute of Cytology and Genetics, Novosibirsk, Russia

Based on publicly available RNA-seq data of human hippocampus samples, we identified alternatively spliced (AS) exons genome wide along with assessing the genes percentage spliced in (psi) values. The data has been compiled on more than 30 samples for each gene. Along with psi values we compiled pairwise covariation matrix across all AS (exon skipping) exons based on Pearson r2. Further Agglomerative Hierarchical Clustering (AHC) procedure on matrix revealed dense AS exons clusters with linked exons, with clusters size in the iterval of [2..19] (pairwise correlation pvalue<1E-6). There were around 2200 genes with clusters of coordinated AS exons. Further analysis revealed AS clusters maintain antagonistic or independent relations within a gene. We explored the traits of genes abundant with AS clusters: the majority proved to be neurospecific genes, including synapse, and cytoskeletal (axonal) genes. Notably, Neurospecific splice factors (SFs) also maintain expanded coordinated AS regulation. While the previous study observed coordinated splicing before [1], the scale of the phenomenon has not been explicitly highlighted. From the evolutionary point of view, and, given the information complexity of splicing decreases upon exon co-varitaion, we may speculate that the rapid response to the homeostatic environment favors quick coordinated splicing mediated tune-up of the gene's isoform. Still, the mechanistic background of phenonenon is not elucidated. One of the viable hypothesis is the specific secondary structure of mRNA favoring the quick coordinated SFs binding. In our report we address the phenomenon and consider several examples of coordinated AS.

Keywords: AS exons co-variation, coordinated alternative splicing within a gene, AS exons cluster, AS mediated information complexity

Acknowledgement: This work has been supported by ICG SB RAS state program.

Sleep Apnea Monitoring using Wearable Heart Rate Sensors

Vladimir Brusić and Yinglun Li

University of Nottingham Ningbo China, Ningbo, China

Sleep apnea is a serious sleep disorder characterized by intermittent stopping and starting of breathing during sleep. It affects approximately 20% of adults globally. Notably, approximately 85% of individuals with moderate to severe sleep apnea remain undiagnosed. Possible complications of sleep apnea include fatigue, high blood pressure, and increased risk of cardiovascular disease. Uncontrolled sleep apnea increases risk or worsens the prognosis of type 2 diabetes, metabolic syndrome, and liver problems.

Sleep apnea diagnosis uses polysomnography (PSG), a monitoring system that records multiple physiological parameters including heart rate, ECG, and oxygen levels during sleep. Although simplified home tests are available, they require multiple concurrent measurements and post-analysis by medical professionals, making them unsuitable for routine home monitoring. The average cost of a sleep apnea test in the USA is approximately $1,100 per test.

We introduce a novel system that utilizes heart rate data from wearable sensors for the monitoring and assessment of sleep apnea suitable for home-based healthcare. Our system utilizes advanced machine learning algorithms to analyze overnight heart rate data collected by wearable sensors. Data analysis is performed using edge computing device. The monitoring system segments the data, employs density maps for feature extraction, deploys machine learning, and assesses the presence and severity of sleep apnea. Finaly, a detailed, clinically relevant report is generated by the system.

The testing of our sleep apnea monitoring system indicates that the accuracy is approximately 90%. Misclassification of existing sleep apnea occurred in patients that have very low heart rate variation. Home-based monitoring systems for sleep apnea can help improve correct diagnosis in affected population and reduce the number of unnecessary tests. Furthermore, the progression of diagnosed sleep apnea can be monitored using a single heart rate sensor.

Keywords: heart rate, home healthcare, machine learning, sleep apnea, wearable sensors

ZEB2 as a driver of human-specific traits: Insights from comparative ChIP-Seq and RNA-Seq

Vladimir M. Jovanović[1,2,*], Jeong-Eun Költzow[1], Amanda Jager Fonseca[1], Sebastian Streblow[1], Katja Ettig[3], Stefano Berto[4] and Katja Nowick[1]

[1] Human Biology and Primate Evolution Group, Freie Universität Berlin, Berlin, Germany
[2] Bioinformatics Solution Center, Freie Universität Berlin, Berlin, Germany
[3] Rudolf-Schönheimer-Institut für Biochemie, Leipzig, Germany
[4] Medical University of South Carolina, Charleston, USA
vladimir.jovanovic@fu-berlin.de

Despite near-identical protein sequences, humans exhibit striking phenotypic differences from other primates. These differences likely stem from subtle changes in gene regulation, not just gene sequences. The transcription factor ZEB2, known for its diverse roles in development and cancer, has emerged as a key player in brain development and neuronal differentiation. We investigated its functional divergence in primates by performing ChIP-Seq with a ZEB2 antibody and RNA-Seq following ZEB2 knockdown in human, chimpanzee, and orangutan B-lymphoblastoid cell lines. Our results showed that ZEB2 binding extends beyond its canonical motif, revealing diverse, potentially species-specific, regulatory preferences. Numerous binding sites within promoter regions exhibited significantly higher affinity in humans, suggesting accelerated evolution of these regulatory elements. While a conserved core of immune-related ZEB2 targets was identified across species, we uncovered 437 potential human-specific targets enriched for chromatin organization and DNA replication functions. Notably, an exceptionally high number of non-coding RNA genes was seen among human-specific targets. Furthermore, ZEB2 knockdown induced a unique pattern of differential gene expression in humans, affecting genes involved in neural development and synaptic organization. This highlights a human-specific functional shift in ZEB2 regulation towards brain-related processes. Our findings not only illuminate ZEB2's potential role in the evolution of uniquely human traits but also provide valuable gene candidates for further functional studies into the genetic basis of our species' distinctive features.

Keywords: human evolution, gene regulation, multi-omics, ZEB2, DNA binding motif

Towards a linearly organised embedding space of biological networks

Alexandros Xenos, Noel-Malod Dognin and Nataša Pržulj*

Barcelona Supercomputing Center, Barcelona, Spain
natasha@bsc.es

Low-dimensional embeddings are a cornerstone in the modelling and analysis of complex biological networks. Embedding biological networks is challenging, as it involves capturing both structural (topological) and semantic information of a graph (i.e., node labels). Typically, nodes with the same label are in the same dense subgraph (neighborhood-based similarity), but it has been shown that similarly annotated nodes can be in different network neighborhoods while having similar wiring patterns (topological similarity). However, current network embedding algorithms do not preserve both types of similarity, which limits the information preserved in the embedding space. Moreover, most existing approaches for mining network embedding spaces rely on computationally intensive machine learning systems to facilitate downstream analysis tasks. On the other hand, word embedding spaces capture semantic relationships linearly, allowing for information retrieval using simple linear operations on word embedding vectors.
In our work, following the NLP paradigm, we introduce novel random-walk-based embeddings that allow mining biological knowledge directly from the embedding space. Namely, we introduce embeddings that locate close in the space genes that have similar biological functions (either topological or neighborhood-based similar nodes). We exploit this property to predict genes participating in protein complexes and to identify cancer-related genes based on the cosine similarities between the vector representations of the genes. We also go beyond embeddings that preserve one type of similarity by introducing novel graphlet-based representations of the networks that simultaneously capture topological and neighborhood membership information. We use all the different network representations to assess whether it is an intrinsic property in the structure of the data (input matrix representation) that yield embedding spaces that enable downstream analysis tasks via simple linear operations. Using nine multi-label biological networks and seven single label networks that are commonly used in machine learning studies, we demonstrate that the more homophilic the network matrix representation, the more linearly organized the corresponding network embedding space, and thus, the better the downstream analysis results. Our results suggest that our new graphlet-based methodologies embed networks into linear spaces, allowing for better mining of the networks and alleviating the need for computational-intensive ML models.

Keywords: bioinformatics, network biology, network embeddings, machine learning

Acknowledgement: This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, Grant Agreement No 860895, the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grant PID2022-141920NB-I00 / AEI /10.13039/501100011033/ FEDER, UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR 01536.

DCS: from Reading Genome to Understanding Life

Xun Xu

BGI Research, Shenzhen, China
xuxun@genomics.cn

The Central Dogma, initially proposed by Francis Crick in 1958, delineates the essential flow of genetic information within living organisms. It succinctly states that DNA directs RNA, which in turn guides protein synthesis. Consequently, understanding the genome is pivotal for comprehending life itself.

DNA Sequencing (D) has emerged as a critical tool in unraveling the mysteries of life. Leveraging cutting-edge platforms like DNB-based high-throughput sequencing, we have amassed vast genomic data. These technological breakthroughs have illuminated various aspects of biology. However, even armed with genomic information, we grapple with a fundamental question: How does the same genomic blueprint yield diverse cell types?

With Single-cell sequencing technology (C), sequencing individual cells, we've gained insights into how the same set of genome orchestrates the intricate dance of cellular forms and functions.

Recent breakthroughs in spatial transcripts (S), particularly techniques like Stereo-seq, to study single cells transcriptome as well as other omics across time and space dimensions. These approaches seek to address critical questions about genome regulation, its impact on cellular diversity, and how these processes influence life phenomena, including aging and disease. The talk will introduce our recent progress of these DCS technologies and application in different biology research projects.

Bioinformatics tools for reconstruction of gene networks of complex diseases

Yuriy L. Orlov[1,3,*], Ekaterina A. Savina[1], Vasilisa A. Turkina[1] and Anastasia A. Anashkina[1,2]

[1] Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia
[2] Engelhardt Institute of Molecular biology RAS, Moscow, Russia
[3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
orlov@d-health.institute

The study of gene networks of complex diseases is an important biomedical task demanding data integration. Despite the existing wide range of computer programs, their adaptation is necessary for clinical data analysis. The adaptation assumes preparation of tutorials, textbooks and training materials for students, interns, and workers of medical institutions without mathematical or computer science background. Such tutorials should be based on available online bioinformatics tools. Here we discuss a scientific project for the study of complex diseases in which it is difficult to identify genetic components, such as cancers, mental disorders, schizophrenia, and Parkinson's disease.
The available software tools have been collected, a data processing pipeline for creating a list of genes associated with given complex disease has been prepared. The list of genes could be compiled based on queries to GEO NCBI (RefSeq), the OMIM (Online Mendelian Inheritance in Man), GeneCards, and MalaCards databases. Then such a list of genes could be refined using other data, such as data on the differential expression of genes (GEO Dataset Browser resource), including non-coding RNAs (from the TCGA database), and information from published papers. The gene ontologies analysis could be performed using open resources for bioinformatic analysis: PANTHER (http://pantherdb.org) and DAVID (https://davidbioinformatics.nih.gov), the g:GOSt resource for visualization of gene ontologies (http://biit.cs.ut.ee/gprofiler/gost).
Next set of tools for gene targets search is related to gene expression. The computer study of sequencing data is based on the integration of available sequencing data (RNA-seq and ChIP-seq) and computer resources: ArrayExpress, TCGA, CCGA, ENCODE (ENCyclopedia Of DNA Elements), as well as local computer resources ANDSystem, TRRD, GeneNet (wwwmgs.bionet.nsc.ru) and ICGenomics (https://www-bionet.sscc.ru/icgenomics/).
As the main examples of applications, the tasks of analyzing brain tumors are considered – for glioma, meningioma, with the study of complications associated with virus infections, including available data published after the coronavirus pandemic. As applications we consider computer reconstruction of gene networks for a number of oncological diseases - glioma, breast cancer, colorectal cancer, and a number of mental disorders such as Parkinson's disease.

Keywords: bioinformatics, gene networks, complex diseases, education, online tools

Acknowledgement: The study was supported by the Russian Science Foundation (grant 24-24-00563).

Impact of Electronic Cigarette Components on Lung Cell Proteome: A High-Resolution Mass Spectrometry Analysis

Aleksandra Divac Rankov[1,*], Sara Trifunović[1], Katarina Smiljanić[2] and Mila Ljujić[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[2] Center of Excellence for Molecular Food Sciences, University of Belgrade - Faculty of Chemistry, Belgrade, Serbia
aleksandradivac@imgge.bg.ac.rs

Electronic cigarettes (e-cigarettes) are a relatively new tobacco product, and still there is much to be determined about the lasting consequences on health of these products. Despite containing mainly non-harmful substances like polyethylene glycol and glycerol, e-cigarettes can also include nicotine and various flavorings, which may contribute to their potential harm.

Our aim was to determine the effects of different components of e-cigarettes on the protein composition of lung cells.

We have performed the comprehensive proteome analysis of epithelial lung cells (BEAS 2B) exposed to e-cigarette vapor condensate. BEAS 2B cells were treated for 24 h with sub-cytotoxic concentration of e-cigarette vapor condensate, made from different e-cigarette liquids - with and without nicotine and with or without flavorings.

The proteome analysis was performed via high resolution mass spectrometry based proteomics (Orbitrap Exploris 240, Thermo Scientific, USA). BEAS 2B proteins were identified with the PEAKS X Pro platform (Bioinformatics Solution Inc., Ontario, Canada) against a UniProtKB database of *Homo sapiens* and contamination database as common Repository of Adventitious Protein entries. All qualitative gene products enrichment analyses have been done with FunRich Software 1.3.1.

We have found reduction in the number of proteins in exposed cells. The most affected cellular components were extracellular exosomes, mitochondria and ribosomes. Presence of nicotine and flavoring together had most effect on the following biological processes: protein translation and proteasome-mediated ubiquitin-dependent protein catabolic processes, tricarboxylic acid cycle and cellular response to interelukin-7.

Our study showed that exposure to e-cigarettes affects critical cellular processes and structures, which could have implications for cell function and overall lung health. Further analysis on pathways identified will help us better understand health risks associated with e-cigarette vaping.

Keywords: electronic cigarette, proteome, lung cells

Acknowledgement: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-66/2024-03/200042)

A pipeline for the identification of disease-specific genetic biomarkers using NGS sequencing data of cfDNAs in human plasma

Alessandra Vittorini Orgeas* and Christoph W. Sensen

FHCEMM, Szeged, Hungary
alessandra.vittorini@hcemm.eu

Circulating cell-free DNAs are DNA fragments released into the blood by different tissues. They can be isolated from a routine blood draw that is a cost-effective, fast and non-invasive procedure. While their presence is detectable under healthy conditions, there are solid evidences of their association with various clinical conditions. This explains the great interest for cell-free DNAs in clinical settings, because of their potential role as biomarkers. Furthermore, the availability of a higher number of samples and high-throughput sequencing technologies have enabled the production of massive amounts of complex genomic data. As consequence, the potentiality of cell-free DNAs can be exploited only if supported by a computational platform that streamlines the execution of the analysis and makes it reproducible and shareable across different platforms. The proposed computational pipeline addresses the complexity of this analysis. The pipeline starts with the raw data pre-processing necessary to remove residual adapters and filter the low-quality reads. It follows the composition analysis where the average sample composition is calculated and expressed in terms of specific target regions of human DNA. Next, a random forest classifier algorithm searches for a subset of the target regions that perform best at predicting the health outcome. The statistical significance of the output is validated by the MANOVA test. Finally, the pipeline runs through the original set of sequencing data to retrieve what sequences best match the composition represented by the pool of target regions deriving from the previous step. The ultimate result is a list of nucleotide sequences that have been identified as the best performing indicators of a clinical condition based on the analysis described above. Thanks to the containerization technology and workflows managers this pipeline can be shared and executed with the same functionalities across different platforms, and its installation process is automated.

Keywords: bioinformatics, computer science, DNA, sequencing, biomarkers

Acknowledgements: We thank Dr. Stefan Grabuschnig (Innophore GmbH, Austria) for the crucial help, guidance and insights in developing the computational framework of the pipeline. This work was funded in part by EU Horizon 2020 Grant No. 739593 and in part by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, with project no. TKP-2021-EGA-05 under the Thematic Excellence Programme and project no. 2022-2.1.1-NL-2022-00005 under the National Laboratory Programme.

Mechanism-based classification of SARS-CoV-2 Variants by Molecular Dynamics Resembles Phylogenetic Tree

Thais Arns[1], Aymeric Fouquier d'Hérouël[1], Patrick May[1], Alexandre Tkatchenko[2] and Alexander Skupin[1,2,3,*]

[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Belvaux, Luxembourg.
[2] Department of Physics and Material Science, University of Luxembourg, Limpertsberg, Luxembourg.
[3] Department of Neuroscience, University of California San Diego, La Jolla, USA
alexander.skupin@uni.lu

The COVID-19 pandemics has demonstrated the vulnerability of our societies to viral infectious disease. The mitigation of COVID-19 was complicated by the emergence of Variants of Concern (VOCs) with varying properties including increased transmissibility and immune evasion. Traditional population sequencing proved to be slow and not conducive for timely action. To tackle this challenge, we introduce the Persistence Score (PS) that assesses the pandemic potential of VOCs based on molecular dynamics of the interactions between the SARS-CoV-2 Receptor Binding Domain (RBD) and the ACE2 residues. Our mechanism-based classification approach successfully grouped VOCs into clinically relevant subgroups with higher sensitivity than classical affinity estimations and allows for risk assessment of hypothetical new VOCs. Interestingly, the PS-based interaction analysis across VOCs resembled the phylogenetic tree of SARS-CoV-2 with high accuracy and reveals for the first time a clear link between sequence determined structures and resulting molecular dynamics further demonstrating the predictive relevance for pandemic preparedness of our approach. Thus, PS allows for early detection of a variant's pandemic potential, and an early risk evaluation for data-driven policymaking.

Keywords: molecular dynamics simulations, SARS-CoV-2 variants, phylogenetic tree, classification (up to 5)

Acknowledgement: This work was supported by the Luxembourg National Research Fund (FNR) COVID-19/21/16874499 – ERCSaCoV

Determining the efficiency of the miTAR neural network in searching for microRNA target genes

A. V. Starostin and D. D. Gavrilova

I.M. Sechenov First Moscow State Medical University, Moscow, Russia
staral.ru@yandex.ru

Introduction: Atherosclerosis is a cardiovascular disease characterized by a chronic inflammatory process of the intima of elastic arteries, which is known to be one of the leading causes of death in developed countries. Recent studies have revealed a correlation between mtDNA mutations and pathological disorders leading to atherosclerosis. Moreover, increasing number of research focuses on the microRNA-mediated gene expression suppression. miRNAs play an important role in the development and regulation of diseases of the cardiovascular system. Consequently, the study of possible linkage between the mitophagy disruption and the miRNA-mediated gene suppression is of interest.
Objective: To assess the effectiveness of neural network based search for target genes for miRNA regulation of atherosclerosis development.
Materials and methods: The list of mitophagy associated genes was selected by analyzing scientific databases. Then, each of the genes was passed through a specialized miRDB database in order to find corresponding miRNA. The resulting miRNA list was analyzed through the web interface, which transmitted data to the miTAR neural network, in order to estimate affinity values. The final stage of the experiment included the identification of the obtained through miTAR sequences through the NCBI Blast web service. Next, the list of obtained genes was compared with the list of genes found manually using miRDB.
Results and discussion: Based on the results of the comparison we came to the conclusion that the neural network based miTAR surpassingly copes with the task of predicting microRNA target genes, albeit the results for miRNA with lower affinity index significantly varied from the ones found manually. However, this reservation could be attributed to upsides of the model as it reveals new potential target genes for further evaluation.
Conclusion: Our results indicate that neural network models can serve as an effective tool in the search for miRNA target genes. In addition, the found genes could be furtherly studied to identify their linkage to atherosclerosis. This will also assist in the study of possible drug substances in the treatment of atherosclerosis.

Leveraging Open Source Hardware and Physics-Informed Machine Learning for Accurate Experimental Identification of Bioink Thermophysical Properties in 3D Bioprinting

Bogdan Kirillov[1,2,*], Katherine Vilinski-Mazur[1] and Dmitry Kolomenskiy[1]

[1] Center of Material Technologies, Skolkovo Institute of Science and Technology, Moscow, Russia

[2] Center for Precision Genome Editing and Genetic Technologies for Biomedicine, Institute of Gene Biology, Russian Academy of Sciences, Moscow, Russia

Bogdan.Kirillov@skoltech.ru

Three-dimensional bioprinting serves as a foundation for a number of modern tissue engineering technologies, for example organ-on-a-chip models and nerve conduits. The ability of 3D bioprinting to create functional tissue structures depends on the properties of the used bioink – the material that consists of cells and supporting structure. Most frequently used supporting component of a bioink for extrusion-based 3D bioprinting is a temperature-activated hydrogel (e.g. kappa carrageenan or sodium alginate), a hydrogel that solidifies when it reaches a specific activation temperature. Use of composite hydrogels that consist of different components (e.g. a combination of hydroxyapatite nanorods and gelatin) allows for precise control of bioprinted construct properties. Knowledge of the hydrogel's activation temperature and the associated thermophysical properties is essential for optimizing the bioprinting process since they influence the settings that one needs to select in order to maximize the probability of successful experiment – printhead temperature, printing speed and extrusion multiplier. Additionally, thermophysical properties have an effect on mechanical properties of the final printed model and also affect the viability of the cells within it. Thermophysical properties, as well as mechanical properties, can be controlled by changing the composition of hydrogel. This study provides a design of experiment for determining thermophysical properties of a composite hydrogel bioink sample using temperature sensors of a 3D bioprinter and a Physics-Informed Machine Learning algorithm. The algorithm combines temperature sensor data, physical simulation of the heat exchange process within a sample and accompanying Machine Learning model that selects the most promising combination of hydrogel components for experimental testing. The experiment is based on a hardware solution of custom open design – the experimental setup can be reproduced using a consumer-grade 3D printer and electronic components readily available on the market. We demonstrate that the combination of open source hardware, artificial intelligence control system and physical simulation allows for accurate assessment of the bioink properties thus making 3D bioprinting more reproducible, robust and accessible.

Keywords: biophysics, 3d bioprinting, thermophysical properties, machine learning, bioinformatics

Analysis of AlphaFold2 Predicted Structures of Aggregation Factors in Lactic Acid Bacteria

Darya Tsibulskaya* and Milan Kojic

Institute of Virology, Vaccines and Sera "Torlak", Belgrade, Serbia
2650460@gmail.com

Autoaggregation is the ability of identical bacterial cells to adhere to each other, which is important for colonization, kin and kind recognition, and bacterial survival.
There is a group of aggregation factors in lactic acid bacteria that have several distinctive features: these are large proteins with a molecular mass greater than 150 kDa, containing an N-terminal signal sequence and an LPXTG-like cell wall anchor domain, as well as a different number of repetitive domains. These aggregation-promoting proteins are also called/known as Snow-flake Forming Collagen Binding Aggregation Factors (SFCBAF) due to their unique aggregation phenotype (PMIDs: 22182285, 29018422, 25955159, 30027759, 38014957). Predictions for different members of this group (predicted by the InterPro program) indicate a varying number and, in some cases, different compositions of repetitive domains.
Comparison of the predicted structures of known aggregation factors (AggL from *Lactococcus lactis*; AggE from *Enterococcus faecium*; AggLb from *Lacticaseibacillus paracasei*; AggLr from *Lactococcus raffinolactis*; and AggA from *Tetragenococcus halophilus*) using AlphaFold2 revealed structural similarities, which may explain the similar phenotype despite low identity (e.g., AggLb is identical to AggA at 21.73% and AggL at 41.14%, while AggA is identical to AggL at 32.43%). The structure itself resembles a shoe-like structure: with a heel and a sole in the form of a loop, consisting of 6-7 adhesion domains superfamily (InterPro: IPR008966). The most structurally dissimilar among them is AggLb, which is the largest of the described aggregation factors of this type and has a greater number of repeats in the second half of the protein compared to other members of this group. The calculation of electrostatic potential shows that the protein surfaces predominantly have negative potential, which is consistent with previously shown data for the AggLb protein, where the strain producing AggLb demonstrated higher affinities for chloroform and a lower percentage of adhesion to ethyl acetate, indicating that AggLb is able to provide strong electron-donor and weaker electron-acceptor features to the bacteria.
In summary, this research enhances our comprehension of the structure of aggregation factors in lactic acid bacteria.

Keywords: AlphaFold2, autoaggregation, SFCBAF.

Acknowledgement: This research was supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, Contract no. 451-03-66/2024-03/200177

The landscape of point mutations leading to pregnancy loss

Evgeniia M. Maksiutenko*, Yury A. Barbitoff,
Yulia A Nasykhova and Andrey S. Glotov

Dpt. of Genomic Medicine, D.O. Ott Research Institute of Obstetrics,
Gynaecology, and Reproductology, St. Petersburg, Russia
evgeniia_maksiutenko@mail.ru

Miscarriage is the most frequent complication of a pregnancy which is devastating for affected families and poses a significant challenge for the health care system. Genetic factors are known to play an important role in the etiology of pregnancy loss; however, despite advances in diagnostics, the causes remain unexplained in more than 30% of cases.

In this work, we aggregated the results of the decade-long studies into the genetic risk factors of sporadic spontaneous abortion and recurrent pregnancy loss (RPL) in euploid pregnancies, focusing on the spectrum of causal point mutations in the fetal genome. A total of 270 genetic variants in 196 unique genes were identified across 31 studies, with the majority of these variants associated with non-recurrent miscarriage. We next aimed to reveal the common properties of the identified genes involved in pregnancy loss. Such an analysis showed that they correspond to broadly expressed, highly evolutionary conserved genes involved in crucial cell differentiation and developmental processes and related signaling pathways. Given these features of known genes, we made an effort to construct a list of 186 candidate genes, variants in which may be expected to contribute to pregnancy loss. We subsequently expanded this work and compiled a database which includes all short genetic variants reported as associated with pregnancy loss. To date, the database includes 479 genetic variants observed in 374 cases, including both euploid fetuses and couples experienced recurrent pregnancy loss.

Taken together, our work summarizes the knowledge about specific genes that contribute to pregnancy loss, and could be of importance in designing a diagnostic sequencing panel for patients and prediction of pregnancy loss risk in couples.

Keywords: pregnancy loss, pathogenic genetic variant, miscarriage, recurrent pregnancy loss

Acknowledgement: This work was supported by the Ministry of Science and Higher Education of Russian Federation (project "Multicenter research bioresource collection "Human Reproductive Health"" contract No. 075-15-2021-1058 from 28 September 2021).

Detecting Genetic Interactions with Visible Neural Networks

Arno van Hilten[1,*], Federico Melograna[2,3,*], Bowen Fan[4],
Wiro Niessen[1,5], Kristel van Steen[2,3,+] and Gennady Roshchupkin[1,6,+]

[1] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands
[2] Department of Human Genetics, KU Leuven, Leuven, Belgium
[3] GIGA-R Molecular and Computational Biology, University of Liège, Liège, Belgium
[4] Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland
[5] Department of Imaging Physics, Delft University of Technology, Delft, The Netherlands
[6] Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
* Authors contributed equally and share first authorship
+ Authors contributed equally and share last authorship
federico.melograna@kuleuven.be

Non-linear interactions among single nucleotide polymorphisms (SNPs), genes, and pathways play an important role in human diseases, but identifying these interactions is a challenging task. Neural networks are state-of-the-art predictors in many domains due to their ability to analyze big data and model complex patterns, including non-linear interactions. In genetics, visible neural networks are gaining popularity as they provide insight into the most important SNPs, genes and pathways for prediction. Visible neural networks use prior knowledge (e.g. gene and pathway annotations) to define the connections between nodes in the network, making them sparse and interpretable. Currently, most of these networks provide measures for the importance of SNPs, genes, and pathways but lack details on the nature of the interactions. Here, we explore different methods to detect non-linear interactions with visible neural networks. We adapt and speed up existing methods, create a comprehensive benchmark with simulated data from GAMETES and EpiGEN, and demonstrate that these methods can extract multiple types of interactions from trained visible neural networks. We also highlight the strengths and weaknesses of the various methods in different settings, providing guidelines for general use-cases. Finally, we apply these methods to a genome-wide case-control study of inflammatory bowel disease and find high consistency of epistasis signals. Follow-up association testing revealed seven statistically significant epistasis SNP pairs. The results and the code to reproduce the analysis are available at https://github.com/ArnovanHilten/GenNet .

Keywords: epistasis, non-linear, interactions, visible, neural networks

Acknowledgement: We would like to acknowledge all the investigators and participants in the International Inflammatory Bowel Disease Genetics Consortium. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreements No 813533 (MLFPM) and No 860895 (TranSYS), the FNRS convention PDR T.0294.24 "Expanded PRS embracing pathways and interactions for increased clinical utility" and through the 2005 Simon Steven Meester grant 2015 to W.J. Niessen by the Dutch Technology Foundation (STW). Work was carried out on the Dutch national e-infrastructure with the support of SURF Cooperative (application number 17610). Gennady V. Roshchupkin supported by the ZonMw Veni grant (Veni 1936320) .

Bioinformatic workflow to analyse Multiome ATAC + Gene Expression data

Iva Sabolić[1,*], Radoslav Atanasoski[2], Robert Šket[3,4], Tine Tesovnik[3,4], Barbara Slapnik[3,4], Klemenitna Črepinšek[3,4], Blaž Vrhovšek[3,4], Tadej Avčin[4,5], Mojca Zajc Avramovič[5], Jernej Kovač[3,4], Uršula Prosenc Zmrzljak[2] and Barbara Jenko Bizjan[2,3,4]

[1] Bioinformatic department, Labena, Zagreb, Croatia
[2] Bioinformatic department, BIA Separations CRO, Labena, Ljubljana, Slovenia
[3] Clinical Institute of Special Laboratory diagnostics, University Children's hospital, University Medical Centre Ljubljana, Slovenia
[4] Faculty of Medicine, University of Ljubljana, Slovenia
[5] Clinical Department of allergology, rheumatology, and clinical immunology, University Children's hospital, University Medical Centre Ljubljana, Slovenia
iva.sabolic@labena.hr

Multi systemic inflammatory syndrome in children (MIS-C) is a rare condition associated with SARS-CoV-2. To improve the understanding of underlying regulatory networks and potentially explain the mechanism behind MIS-C disease onset, we conducted a simultaneous profiling of transcriptome and epigenome from the same cell using the 10X Genomics Chromium Single Cell Multiome ATAC + Gene Expression protocol. Our study design consisted of 10 patients sampled at two time points: at the time of MIS-C disease flare before treatment was applied, and at the time of disease remission.

In order to facilitate an efficient way of processing the wealth of data generated by the employed protocol, we developed a bioinformatic workflow that enables users to systematically extract information crucial for understanding cellular function and regulation. The workflow is comprised of primary analysis performed using 10X Genomics Cell Ranger ARC software, and advanced analysis conducted utilizing a curated collection of various R packages. Through the proposed analyses, researchers can effectively identify and correct technical aberrations and batch effects in the data, unveil distinct cell types, detect differential gene expression between flare and remission states, reveal heterogeneity within cell populations, pinpoint enriched biological pathways and functions, and elucidate regulatory elements controlling gene activity.

The results of this project are expected to enlighten the underlying pathophysiology of MIS-C flare and support clinical decision on more targeted treatment. The identified disrupted networks during MIS-C flare could lead the way to establish an early diagnosis and improve long-term outcome, including prevention of myocardial and neuropsychological impairment.

Keywords: bioinformatic pipeline, multiome, COVID-19, MIS-C

Acknowledgement: Grants received form: ARIS: J3-50115, J3-3061, Interreg Italia-Slovenia: Concerto

Application of Kolmogorov-Arnold Networks in Cervical Cancer Diagnostics

Ivan Lorencin[1,*], Nikola Tanković[1], Ariana Lorencin[2] and Matko Glučina[3]

[1] Juraj Dobrila University of Pula, Faculty of Informatics, Pula, Croatia
[2] Department of Gynecology and Obstetrics, General Hospital Pula, Pula, Croatia
[3] Istrian University of Applied Sciences, Pula, Croatia
ivan.lorencin@unipu.hr

In this study, we explore the application of Kolmogorov-Arnold Networks (KAN) in the field of cervical cancer diagnostics. Cervical cancer, a major health concern worldwide, requires efficient and accurate diagnostic methods for early detection and treatment. Traditional machine learning approaches, such as multilayer perceptron (MLP) and K-nearest neighbors (KNN), have shown high classification performance but often at the cost of complex and resource-intensive architectures. In contrast, KAN offers a simpler yet effective alternative.

Utilizing a publicly available dataset of cervical cancer data, which includes 859 samples with 36 input attributes and diagnostic outputs defined as Hinselmann, Schiller, cytology, and biopsy, we implemented KAN for classification. Given the significant class imbalance in the dataset, we also applied various class balancing techniques.

Our results indicate that KAN can achieve high classification performance with mean area under the receiver operating characteristic curve (AUC) and mean Matthew's correlation coefficient (MCC) scores comparable to those obtained with more complex architectures. Specifically, the KAN models demonstrated robust diagnostic capabilities, achieving AUC and MCC scores above 0.9.

The simplicity of KAN architectures, combined with their strong performance metrics, underscores their potential as a practical tool in medical diagnostics. These findings suggest that Kolmogorov-Arnold Networks could be effectively utilized for cervical cancer screening and diagnosis, providing a balance of high accuracy, robustness, and reduced computational complexity. This approach could facilitate more accessible and efficient diagnostic processes, particularly in resource-limited settings.

Keywords: cervical cancer, class balancing techniques, Kolmogorov-Arnolds network, preventive screening,

Efficient Large Scale Multimodal Image Registration

Joakim Lindblad

Centre for Image Analysis, Department of Information Technology,
Uppsala University, Uppsala, Sweden
joakim.lindblad@it.uu.se

Multimodal imaging refers to the capturing of complementary information about a specimen by different imaging techniques (modalities). Such complementary information allows reaching deeper understanding and improved analysis and diagnostics performance. Multimodal imaging combined with correlated analysis of the acquired data can be very useful for both human and AI-based decision making. For successful correlation and fusion of the heterogeneous information, acquired images need to be accurately aligned – a task which is far from easy, given the great diversity of imaging modalities and specimens, combined with the typically very large size of medical and biomedical images.
In this work we present a computationally efficient method that reaches a state-of-art performance for multimodal image registration. The method is based on computing the cross-mutual information function (CMIF) through efficient evaluation of mutual information in the Fourier domain for every possible discrete translation. Utilizing the power of GPU-based processing and performing a search over a limited set of rotation angles, the approach facilitates accurate rigid alignment at high speed.
We demonstrate how this approach can be used for improved deep learning-driven oral cancer detection by practically enabling information fusion from different imaging modalities on whole slide image data, overcoming problems originating from stitching artefacts and microscope drift.

Keywords: mutual information, image alignment, correlated imaging, whole slide imaging, cytology

Acknowledgement: We are grateful for the scientific support of J. Öfverstedt and N. Sladoje. The work was financially supported by the Swedish research council (grants 2017-04385 and 2022-03580), Sweden's Innovation Agency (VINNOVA) (grants 2017-02447, 2020-03611, 2021-01420), and Cancerfonden (grants 22 2353 Pj and 22 2357 Pj).

MONFIT: Multi-omics factorization-based integration of time-series data sheds light on Parkinson's disease

Katarina Mihajlović[1,*], Noël Malod-Dognin[1], Corrado Ameli[2], Alexander Skupin[2,3,4] and Nataša Pržulj[1,5,6]

[1] Barcelona Supercomputing Center (BSC), Barcelona, Spain
[2] The Integrative Cell Signalling Group, Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Esch-sur-Alzette, Luxembourg
[3] Luxembourg Institute of Health (LIH), Esch-sur-Alzette, Luxembourg
[4] University of California San Diego, La Jolla, CA 92093, USA
[5] Department of Computer Science, University College London, WC1E 6BT London, United Kingdom
[6] ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Spain
katarina.mihajlovic@bsc.es

Parkinson's disease (PD) is a severe and complex multifactorial neurodegenerative disease whose elusive pathophysiology prevents the development of curative treatments. Studying PD using longitudinal multi-omics data is a promising approach to identifying its mechanisms of etiology and progression. However, heterogeneous data require new analysis frameworks that can utilize the complementary information captured by diverse data types and further the understanding of PD across biological entities and processes.

We present MONFIT, a holistic analysis pipeline that integrates and mines time-series single-cell RNA-sequencing data of disease and control cell lines, along with bulk proteomics and metabolomics data, by non-negative matrix tri-factorization, hence enabling prior knowledge integration from molecular networks. MONIFT first integrates (fuses) time-point-specific data, producing time-point-specific gene embeddings, which it then collectively mines across time points.

We apply MONFIT to longitudinal, multi-omics data of PD and control cells obtained from patient-derived induced pluripotent stem cells that were differentiated into dopaminergic neurons. We predict 123 genes related to PD, which we validate by network analysis to be specific to the PINK1 mutation causing PD. We investigate the top 30 gene predictions and propose five novel PD gene candidates: CENPF, CRABP1, TOP2A, TMSB10, and NASP. In addition, we emphasize molecular pathways that play important roles in PD pathology and suggest new intervention opportunities by drug repurposing. We demonstrate that MONFIT goes beyond standard differential analysis approaches of single-omics data by predicting PD-associated genes that would otherwise elude discovery. MONFIT is a generic method and can be modified to accommodate data from tissue samples and other multi-omics data types.

Keywords: data fusion, data mining, single-cell data, network biology, multi-omics data

Acknowledgement: This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, Grant Agreement No 860895, the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grant PID2022-141920NB-I00 / AEI /10.13039/501100011033/ FEDER, UE, and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR 01536

PCR-based nanopore sequencing method for characterizing short tandem repeat expansions

Lana Radenkovic[1,*], Jovan Pesovic[1], Vladimir Tomic[2], Igor Davidovic[1], Nemanja Radovanovic[1], Ana Popic[2] and Dusanka Savic-Pavicevic[1]

[1] University of Belgrade - Faculty of Biology, Belgrade, Serbia
[2] Centogene, Belgrade, Serbia
lana.radenkovic@bio.bg.ac.rs

Short tandem repeat (STR) expansions cause >60 rare neurological diseases and pose a methodological challenge due to their length (0.15->3 kb), complex sequence structure, and a continuous increase in length in somatic cells during a patient's life. These genetic features are the main source of extreme variability in the presentation and progression of the diseases. To make genetic diagnosis more accessible and improve prognosis in individual patients, we are developing a nanopore sequencing method to characterize STR expansions deeply.

The method is based on PCR enrichment and a bioinformatics pipeline developed in-house. The library containing up to six patient samples was prepared using Native Barcoding Kit 24 V14. Sequencing was performed on R10.4.1 Flongle cells on a portable Mk1C device (Oxford Nanopore Technologies). Raw data were basecalled with Guppy. Reads were analyzed as complex strings with an STR between predefined locus-specific flanking sequences. Regular expressions were used to determine the length and structure of the STRs. Reads were categorized based on the STR length and aligned without reference. Polishing steps for overcoming per-read errors were based on results from our wet lab experiments.

The length and structure of STR expansion in patients with spinocerebellar ataxia type 8 (e.g., $(CTA)_6(CTG)_{56-60}CCG(CTG)_{53-56}$) and myotonic dystrophy type 1 (DM1) (e.g., $(CTG)_{350-700}(CCGCTG)_3(CTG)_4(CCGCTG)_2CTGCCG(CTG)_{18}$) were accurately determined compared to orthogonal methods. For somatically highly unstable DM1 mutation, a reliable distribution of allele length was achieved with at least 200X coverage, allowing accurate estimation of modal allele size and degree of somatic instability.

Our nanopore sequencing method can reliably characterize clinically relevant features of STR expansions: length, sequence structure and degree of somatic instability. The method outperforms gold standard methods (repeat-primed PCR, Southern blot), captures somatic variability more reliably compared to Cas9-based enrichment and is more accessible for clinical settings compared to PacBio and Illumina (where applicable) sequencing methods.

Keywords: STR expansions, nanopore sequencing, regular expressions, long-read sequencing

Acknowledgement: This research was supported by the Science Fund of the Republic of Serbia, Grant number 7754217, Understanding repeat expansion dynamics and phenotype variability in myotonic dystrophy type 1 through human studies, nanopore sequencing and cell models – READ-DM1.

Intrinsic disorder of proteins associated with diseases

Lazar Vasović* and Jovana Kovačević

Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
pd212006@alas.matf.bg.ac.rs

Numerous publicly accessible databases include variously formatted information regarding the relationship between genes and diseases. This work expedites their use by integrating them into one standardised database – Integrated Gene Disease Database. IGDD currently has more than 400,000 rows incorporating gene-disease associations from the following sources: DisGeNet, COSMIC, HumsaVar, Orphanet, ClinVar, HPO, DISEASES. Its features include: gene symbol and IDs, UniProt ID, disease name, Disease Ontology ID. Disease Ontology was chosen since it offers a wide range of possibilities in terms of disease exploration.

IGDD was further enriched with information on the disorder of the proteins encoded by the genes associated with diseases since many lack a fixed and well-defined three-dimensional structure. That fact may be linked with the disease-causing mechanisms, so it is an important feature of a protein. Several disorder measures were used, based both on the sequence profiling and the advanced statistical methods: amino acid profiles, charge-hydropathy (CH) prediction, PONDR family (VL-XT, VSL2), IUPred family (long, short, ANCHOR), FuzDrop.

This work focuses on the following question: is there any relationship between certain diseases or their groups and the level of disorder of proteins related to them? With that in mind, no correlation was found between any considered disorder measure and the number of diseases that proteins are related to. There was neither a correlation between the depth of diseases in the ontology and the disorder of the related proteins. Additionally, no obvious regularity was noticed when it comes to the disorder of proteins grouped by diseases they are related to. Both ordered and disordered proteins were equally found in all parts of the ontology.

Regardless of the results of this research, IGDD can nevertheless be considered a valuable resource for future data analysis and further investigation of gene-disease associations. Its detailed features and a large number of relations open the path for many types of studies.

Keywords: gene-disease associations, protein disorder, correlation analysis

Acknowledgement: This research was financially 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 (Lazar Vasović, contract number 451-03-1271/2022-14/2990) and through Project No. 174021. It is based upon work from COST Action CA21160, named Non-globular proteins in the era of Machine Learning (ML4NGP) and supported by COST (European Cooperation in Science and Technology).

Label-Free Quantitative Proteomics of *Pelargonium zonale*: Tissue-Specific Differences

Dejana Milić[1], Thierry Balliau[2], Marlène Davanture[2], Melisande Blein-Nicolas[2] and Marija Vidović[1,*]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
[2] Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE–Le Moulon, PAPPSO, France
mvidovic@imgge.bg.ac.rs

Variegated *Pelargonium zonale* plants present excellent model for studying metabolite fluxes and photosynthesis-related processes between the photosynthetically active (green, vG) and inactive (white, vW) tissues within the same leaf under the same microenvironmental conditions. We aimed to investigate the proteomic differences between these two metabolically contrasting tissue types, as well as between vG and plain morphs (G) to gain more insight into the evolutionary benefits of variegation.
Label-free proteomics was performed using liquid chromatography coupled to tandem mass spectrometry. The data were searched against the newly expanded RNAseq database for *P. zonale* leaves. Analysis was done with X!Tandem, using 10 ppm precursor mass precision and fragment 0.02 Th mass tolerance. The identified proteins were filtered using X!TandemPipeline, requiring at least two peptides with E-values lower than 0.01 and a protein E-value<10^{-5}. Peptide ions, and their parental proteins, were quantified by integrating signal intensities from extracted ion currents (XIC) using MassChroQ software. A total of 2707 protein groups were identified. After removing dubious data and peptides with missing values, we obtained 2009 protein group. We annotated 564 and 79 differentially abundant proteins in vG vs. vW and vG vs. G, respectively. The differentially abundant proteins were mapped into metabolic pathways using the modified MapMan 3.6.0RC1 software. Proteins related to photosynthesis and carbohydrate metabolism were more abundant in vG compared with vW, while proteins related to oxidative stress and protein degradation were more abundant in vW than in vG. Compared to vG, G tissue contained more proteins involved in energy production and protein synthesis. Briefly, this study has paved the way to uncover the evolutionary advantages of the variegated phenotype.

Keywords: differential proteomic analysis, variegated *Pelargonium zonale*, metabolic pathways

Acknowledgement: This work was funded by the Ministry of Science, Technological Development and Innovation, Republic of Serbia (Contract No. 451-03-47/2024-01/200042, 2022), Bilateral project with Republic of France (no. 337-00-93/2023-05/3).

Interaction between healthy and diseased bronchial epithelial cells with *Lactiplantibacillus plantarum* BGPKM22 reveals distinct expression profiles by dual RNA sequencing

Marija Stankovic*, Hristina Mitrovic, Svetlana Sokovic-Bajic, Katarina Veljovic and Natasa Golic

Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
marijast@imgge.bg.ac.rs

Although low density features lung microbiota, its composition is crucial for the development and maintenance of healthy human lungs. In chronic obstructive pulmonary disease (COPD) lung microbiota is changed, correlating with the disease severity and exacerbations. But whether changed lung microbiota causes COPD, or microbiota changes due to disease is unclear. Moreover, the effect of COPD on microbes is unknown.
We aimed to scrutinize the effects of interspecies interaction between healthy and diseased human bronchial epithelial cells with bacteria featured by preferred properties, *Lactiplantibacillus plantarum* BGPKM22, by dual RNA sequencing.
Primary healthy and diseased bronchial epithelial cells, from a healthy subject and COPD patient, respectively, were exposed to *Lactip. plantarum* BGPKM22. Dual RNA sequencing and data processing were performed by Novogene (Beijing, China). Clean reads were aligned to the human genome and *Lactip. plantarum* SK151 using Hisat2 tool. Differential gene expression analysis was performed using the edgeR R package. The adherence of BGPKM22 to the cells and its resistance to oxidative stress were determined.
In healthy and diseased cells interaction with BGPKM22 caused a change in expression of 52 and 45 genes, respectively. The genes *IQCN*, *LINC01554*, *KCNB1*, and *CDK7* indicated a specific response of human bronchial epithelial cells exposed to BGPKM22, regardless of the health status. Markedly more genes showed a change in expression in BGPKM22 in interaction with healthy than with diseased cells, 486 and 101, respectively. The adhesion of BGPKM22 was better to healthy than to diseased cells. The fitness of BGPKM22 increased only after interaction with healthy cells.
Utilization of BGPKM22 can alleviate symptoms and replenish diminished lung microbiota in COPD. Preferential affinity of BGPKM22 towards healthy cells can explain diminished lung microbiota in COPD. Beneficial effects of BGPKM22 can remain unexploited by host due to decreased affinity and fitness of BGPKM22 in interaction with diseased cells. Abundant response of BGPKM22 in interaction with healthy cells, sheds a new light on potential lung probiotics depending on the host state.

Keywords: primary human bronchial epithelial cells, *Lactiplantibacillus plantarum* BGPKM22, interspecies interaction, dual RNA sequencing.

Acknowledgement: This research was funded by the Science Fund of the Republic of Serbia, grant PROMIS, #6066974 LABLUNG.

Impact of Protein Representations on Drug-Target Affinity Prediction

Matija Marijan[1, 2,*] and Ivan Tanasijević[1]

[1] The Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
[2] School of Electrical Engineering, University of Belgrade, Belgrade, Serbia
matija.marijan@ivi.ac.rs

Accurate and rapid prediction of the binding affinity between potential drug candidates and target proteins can significantly hasten the drug discovery and development process. Utilizing artificial intelligence (AI) models to predict drug-target affinity (DTA) is an affordable and efficient strategy for sifting out undesirable molecules and identifying promising drug candidates. This approach allows researchers to focus on the most promising compounds for further in silico and wet lab experiments, thereby streamlining the overall workflow.
Advancements in AI research, such as the development and implementation of graph neural networks (GNN) and attention mechanisms, have significantly improved methods for processing small molecules as potential drug candidates. These developments now allow for very efficient and accurate DTA prediction, without the need for extensive protein processing resources. While this progress marks a significant step forward in computational drug discovery, models that heavily rely on efficient molecule processing may still lack the incorporation of highly specific protein information into their algorithms, which could be crucial for further improvement.
In this study, we present a comprehensive analysis of the impact of different protein representations on the accuracy of DTA prediction using two datasets, by implementing and modifying AI models that are based on GNNs and large language models (LLM).
Motivated by the intuitive resemblance between traditional motif search methods for protein sequence analysis and conventional one-dimensional convolution in AI signal processing, we propose a protein representation model based on transposed convolutional neural network (NN) layers. Preliminary results indicate that such embeddings improve the overall affinity prediction accuracy, compared to similar models from the literature. Additionally, implementing LLMs to generate protein embeddings independently of other NN layers has demonstrated potential to significantly enhance the accuracy of predicting drug-target pairs that have a very low or unmeasurable affinity.

Keywords: drug-target affinity, protein embeddings, graph neural networks, large language models, motif search

Estimating the dimensionality of omics network embedding space

Milena Stojic[1], Noël Malod-Dognin[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
natasha@bsc.es

Thanks to the advances in capturing technology, huge amounts of large-scale biological, omics data have been accumulated. These data are naturally modeled as networks in which nodes represent entities (e.g., patients, genes, metabolites) and edges represent interactions between them. Because of the computational complexity of directly mining networks, current approaches first embed these networks in low-dimensional vector space, and then mine the resulting node embedding vectors for new biomedical knowledge. However, despite successful applications of network-embedding methodologies for mining biological data, there is still no gold-standard approach for determining its key parameter; the number of dimensions of the embedding space. Thus, to set this parameter, most studies rely on computationally inefficient grid-searches. Recently, Two Nearest Neighbors (2NN), a methodology that estimates the intrinsic dimensionality of data-points in high dimensional space, has been successfully applied to estimate the number of dimensions needed to embed synthetic and toy example networks.
In this work, we investigate the applicability of 2NN for determining the dimensionality of biological, omics network embedding spaces. On the protein-protein interaction networks and the gene co-expression networks of budding yeast and of homo sapiens, we relate the obtained dimensionality estimations with various network topological properties and with biomedical downstream analysis tasks.

Keywords: bioinformatics, network data mining, network embedding, network biology, AI

Acknowledgement: This project has received funding from the European Union's EU Framework Programme for Research and Innovation Horizon 2020, Grant Agreement No 860895, the European Research Council (ERC) Consolidator Grant 770827, the Spanish State Research Agency and the Ministry of Science and Innovation MCIN grants PID2022-141920NB-I00/AEI/10.13039/501100011033/FEDER, UE, and PID2022-141920NB-00/AEI/10.13039/501100011033/ERDF, UE Project: PN046500 and the Department of Research and Universities of the Generalitat de Catalunya code 2021 SGR 01536.

Enhancing Biomedical Information Retrieval with Semantic Search: A Comparative Analysis Using PubMed Data

Adela Ljajić[1], Lorenzo Cassano[2], Miloš Košprdić[1,*],
Bojana Bašaragin[1], Darija Medvecki[1] and Nikola Milošević[2]

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

PubMed excels in retrieving scientific articles through keyword matching in biomedical literature. However, its efficacy in comprehending and addressing natural language queries is limited due to its emphasis on basic text matching and absence of contextual understanding. This limitation becomes challenging when users pose inquiries in natural language that do not align with the structured vocabulary of the database. To address this, we are presenting an Information Retrieval System utilizing indexed data sourced from PubMed articles (title+abstract), which employs a combination of lexical and semantic search to retrieve the most accurate responses to user inquiries.

For the vector representation of concatenated titles and abstracts, we employed a sentence transformer model optimized for asymmetric semantic search, given our focus on shorter queries searching through longer texts. Lexical indexing utilized the OpenSearch database, while semantic indexing was facilitated by Qdrant. Tuning the hybrid search results achieved an optimal balance between lexical and semantic search parameters. Evaluation was conducted using the BioASQ dataset comprising 5049 questions, each paired with PubMed articles and annotated by domain experts. We also used this dataset to assess the performance of the PubMed Search Engine in biomedical question answering, enabling a comparative analysis.

Utilizing the lexical index for document retrieval yielded MAP@10 of 0.411. Through experimentation, we determined that the optimal hybrid query combination entails weights of 0.7 and 0.3 for lexical and semantic components, respectively. Integrating the best lexical results with the semantic index led to an enhanced MAP@10 of 0.425. Assessment of the PubMed search engine on the same BioASQ dataset unveiled MAP@10 of 0.153 when MeSH terms were omitted and MAP@10 of 0.191 when they were included in the search of the PubMed database. Our system notably advances biomedical information retrieval by leveraging a fusion of lexical and semantic search, resulting in heightened precision when responding to natural language queries, surpassing PubMed's keyword-based approach.

Keywords: PubMed, information retrieval, vector database, LLM's, hybrid search

Acknowledgement: The project Verif.ai is a collaborative effort of Bayer A.G. and the Institute for Artificial Intelligence Research and Development of Serbia, funded within the framework of the NGI Search project under Horizon Europe grant agreement No 101069364.

STIM: Multipurpose method for spatial transcriptomics data integration across different technologies

Milos M. Radonjic[1,*], Aleksandra Stanojevic[1], Tamara Banovac[1], Fang Shuangsang[2,3] and Junhua Li[1,3]

[1] BGI Research, Belgrade 11000, Serbia
[2] BGI Research, Beijing 102601, China
[3] BGI Research, Shenzhen 518083, China
milosradonjic1@genomics.cn

We developed an innovative, statistically based data integration method specifically tailored for spatial transcriptomics data. Our method successfully performed all data integration tasks, while removing batch effects by correcting the entire gene expression matrix, ensuring superior preservation of biological information. The outstanding preservation of biological information is significantly enhanced by employing piece-wise affine transformations for aligning gene expression distributions across samples. Our technique robustly demonstrated exceptional batch-effects correction performances across various experimental technologies, datasets, and integration tasks by outperforming all existing methods, especially in preservation of biological information.
As an integral feature, we developed an entirely new spatially aware clustering method capable of accurate identification of tissues and spatial domains. Together with a novel cross-sample clusters mapping methodology, the method ensured robust cross-sample clustering applicable to spatial domains clusters, as well as cell-type clusters. Due to all these features, our method demonstrated remarkable versatility, enabling batch-effects-free integration of multiple samples, 3D clustering, and the seamless incorporation of healthy and diseased samples.
Moreover, our method is the one and only that directly corrects a gene expression matrix by applying transformations which keep the gene regulatory information preserved, allowing studying of gene regulations and gene co-expressions after data integration. This makes our method unique and the only choice for any downstream analysis task which requires full gene expression matrix as an input.

Keywords: bioinformatics, spatial transcriptomics, data integration, batch effects

Detecting somatic copy number variations in 245,388 participants from All of Us biobank

Milovan Suvakov[1], Zhiyv Niu[2] and Alexej Abyzov[1]

[1] Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, Minnesota, USA
[2] Department Laboratory Medicine and Pathology, Mayo Clinic College of Medicine, Rochester, Minnesota, USA
suvakov.milovan@mayo.edu

In recent years, the study of mosaic mutations in human tissues gain significant attention due to advancement in methodology and late data biobanks. As part of that clonal hekotapaisis (CHIP) and its implications in age-related diseases, has attracted considerable attention. CHIP, a prevalent phenomenon in aging individuals, is linked with all-cause mortality, blood cancer, and cardiovascular disease risks, but also exhibits protective effects against conditions like Alzheimer's disease.
We have developed a new methodology implemented in CNVpytor, that detects mosaic copy number variation (mCNV) from WGS by leveraging two independent signals from sequencing data: (1) depth of mapped reads; (2) B-allele frequency of SNPs and small indels. This technique allows for the detection of somatic mCNVs, down to 1% cell frequency. To improve quality of our detection we considered evidence from discordant read pairs and SNP genotyping array data.
Our initial analysis of data from 245,388 individuals in the All of Us (AoU) cohort led to the identification of 2,607 large (>10Mbp) confident somatic mCNVs. We observed an expected trend where older individuals exhibited a higher number of somatic CNAs, consistent with the understanding that detectable clonal hematopoiesis increases with age. Our investigation into chromosome Y loss (LOY) among male samples revealed that over 20% exhibit LOY, indicating a higher prevalence than other somatic mCNVs. Additionally, we found hundreds of thousands smaller mCNVs. The discovery of a small mCNVs in a young individuals, presumed to have originated during development, indicates that analysis of all AoU samples can be useful for understanding of the differences in the occurrence and nature of CNAs during development compared to those in aging.
This comprehensive analysis is expected to result in a shared computational resource, offering mCNV calls for the wider research community. By providing these resources, we aim to not only augment the value of AoU data but also establish a foundation for future research methodologies as the AoU's sample collection expands.

Keywords: clonal hematopoiesis, mosaic copy number variation, copy number alterations, somatic mutations, aging

Acknowledgement: This work is supported by the National Institutes of Health (grant no 1R03AG085705) and Mayo Clinic DLMP Scholarly Clinician Award.

Z-Flipons conserved between human and mouse are associated with increased transcription initiation rates

Nazar Beknazarov

Higher School of Economics, Moscow, Russia
halonazar1997@gmail.com

A long-standing question concerns the role of Z-DNA in transcription. Here we use a deep learning approach based on the published DeepZ algorithm that predicts Z-flipons based on DNA sequence, structural properties of nucleotides and omics data. We examined Z-flipons that are conserved between human and mouse genomes after generating whole-genome Z-flipons maps by training DeepZ on ChIP-seq Z-DNA data, then overlapping the results with a common set of omics data features. We revealed similar pattern of transcription factors and histone marks associated with conserved Z-flipons, showing enrichment for transcription regulation coupled with chromatin organization. 15% and 7% of conserved Z-flipons fell in alternative and bidirectional promoters. We found that conserved Z-flipons in CpG-promoters are associated with increased transcription initiation rates. Our findings empower further experimental explorations to examine how the flip to Z-DNA alters the readout of genetic information by facilitating the transition of one epigenetic state to another.

A novel approach to SARS CoV-2 classification

Biljana T. Stojanović[1], Saša N. Malkov[2], Miloš V. Beljanski[3], Gordana M. Pavlović Lažetić[2], Mirjana M. Maljković Ružičić[2], Ivan Lj. Čukić[2] and Nenad S. Mitić[2,*]

[1] Mathematical Institute SASA, Belgrade, Serbia,
[2] University of Belgrade, Faculty of Mathematics, Belgrade, Serbia
[3] Institute for General and Physical Chemistry, Belgrade, Serbia
nenad.mitic@matf.bg.ac.rs

This paper presents an approach for clustering of particular SARS-CoV-2 protein types based on Codon Usage (CU) bias measures. Our previous research has shown that clustering based on CU bias measures is very close to the natural clustering by protein type, regardless of virus affiliation. Relative Synonymous Codon Usage, RSCU, Effective Number of Codons, ENC along with Effective Number of Codons for individual AAc, ENCAA and Relative Codon Bias Strength, RCBS were calculated to measure the CU bias in different proteins coding sequences.
The dataset contains 928.850 SARS-CoV-2 complete virus isolates with non-ambiguous nucleotide sequences. It contains 1.145.168 unique (out of a total of 15.564.504) protein nucleotide sequences and the corresponding AAc sequences. Protein coding sequences are associated with metadata, including the collection date and the WHO virus strain annotation.
Protein coding sequences within the same type (for each of the 12 most abundant types) were clustered. Different clustering algorithms (BIRCH, Kohonen Neural Network, fuzzy and probabilistic clustering) were performed for clustering proteins based on RSCU, ENC and RCBS with a variable number of clusters. WHO group annotations were used for additional cluster description. Most clusters in all results are homogeneous (with a maximum size of about 19-35% of the input material) and are almost pure related to specific WHO group. Each result contains one or two small cardinality heterogeneous clusters with mixed WHO groups. These heterogeneous clusters likely denotes proteins (isolates) that were present at the transition between the two WHO groups. Combining results from different clustering algorithms the membership to WHO groups of SARS-CoV-2 proteins can be described with very high accuracy using protein clustering based on the results of CU bias measures.

Keywords: SARS-CoV-2 WHO groups, codon usage, clustering, data mining

Deciphering the effects of nanosized polystyrene particles using lab-on-chip technology and transcriptome profile

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

[1] Institute for Information Technologies, University of Kragujevac, Kragujevac, Serbia
[2] Faculty of Science, University of Kragujevac, Kragujevac, Serbia
[3] Faculty of Medical Sciences, University of Kragujevac, Kragujevac, Serbia
[4] BIA Separations CRO Laboratory, Ljubljana, Slovenia
[5] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[6] Labena d.o.o. Serbia, Belgrade, Serbia
[7] BioIRC - Bioengineering Research and Development Center, University of Kragujevac, Kragujevac, Serbia
[8] Faculty of Engineering, University of Kragujevac, Kragujevac, Serbia
nevena.milivojevic@uni.kg.ac.rs

Transcriptome profiling at the single cell level is crucial for understanding complex biological systems and molecular mechanisms. We wanted to unravel the influence of polystyrene nanoparticles on peripheral blood mononuclear cells (PBMCs), using microfluidic technology. A total of 4 single-cell sequencing libraries were analyzed (one control and three different treatments). Thousands of individual cells per sample are Barcoded separately to index the transcriptome of each cell individually. Raw sequencing data were analyzed with the Cell Ranger software and visualized using Loupe Browser software. Set of analysis pipelines processes Chromium Single Gene Expression data to align reads, generate Feature Barcode matrices, and perform clustering and gene expression analysis. Each element of the matrix is the number of UMIs (Unique Molecular Identifier) associated with a feature (row) and a barcode (column). Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) algorithms on single-cell sequencing samples were carried out. The expressed cells were clustered during which typical cell marker genes were used for annotation. Genes showing adjusted p-value < 0.05 and |log2 (fold change) | > 0.5 were considered to be marker genes. Loupe Browser was used for visualization of clusters and analysis of the single-cell data. In this way, gene markers for individual cell types obtained by single-cell sequencing represent a good model for the analysis of biological events.

Keywords: bioinformatics, single cell, sequencing, scRNA-seq, microfluidic

Acknowledgement: This research is funded by Labena Slovenia 10xGenomics Grant Challenge. This research was supported by Labena Serbia, the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, contract number 451-03-66/2024-03/200378 (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.

Epigenome-wide analysis identifies a methylome profile linked to Obsessive-Compulsive Disorder, disease severity, and treatment response

Rafael Campos-Martin[1,*], Katharina Bey[2,3], Björn Elsner[5], Benedikt Reuter[5,6], Julia Klawohn[5,6], Norbert Kathmann[5], Michalel Wagner[2,3,4] and Alfredo Ramirez[1,3,4]

[1] Division of Neurogenetics and Molecular Psychiatry, Department of Psychiatry and Psychotherapy, University of Cologne, Medical Faculty, 50937 Cologne, Germany
[2] Department of Psychiatry and Psychotherapy, University Hospital Bonn, Bonn, Germany
[3] German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[4] Department for Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany.
[5] Department of Clinical Psychology, Humboldt-Universität zu Berlin, Berlin, Germany
[6] Department of Medicine, Medical School Berlin MSB, Berlin, Germany
rafael.campos-martin@uk-koeln.de

Obsessive-compulsive disorder (OCD) is a prevalent mental disorder affecting ~2–3% of the population. This disorder involves genetic and, possibly, epigenetic risk factors. The dynamic nature of epigenetics also presents a promising avenue for identifying biomarkers associated with symptom severity, clinical progression, and treatment response in OCD. We, therefore, conducted a comprehensive case-control investigation using Illumina MethylationEPIC BeadChip, encompassing 185 OCD patients and 199 controls recruited from two distinct sites in Germany. Rigorous clinical assessments were performed by trained raters employing the Structured Clinical Interview for DSM-IV (SCID-I). We performed a robust two-step epigenome-wide association study that led to the identification of 305 differentially methylated CpG positions. Next, we validated these findings by pinpointing the optimal set of CpGs that could effectively classify individuals into their respective groups. This approach identified a subset comprising 12 CpGs that overlapped with the 305 CpGs identified in our EWAS. These 12 CpGs are close to or in genes associated with the *sweet-compulsive brain hypothesis* which proposes that aberrant dopaminergic transmission in the striatum may impair insulin signaling sensitivity among OCD patients. We replicated three of the 12 CpGs signals from a recent independent study conducted on the Han Chinese population, underscoring also the cross-cultural relevance of our findings. In conclusion, our study further supports the involvement of epigenetic mechanisms in the pathogenesis of OCD. By elucidating the underlying molecular alterations associated with OCD, our study contributes to advancing our understanding of this complex disorder and may ultimately improve clinical outcomes for affected individuals.

Keywords: data mining, psychiatry, epigenetics, Obsessive-Compulsive Disorder

Unraveling Bacterial Persister Formation: Insights from a Type I Toxin-Antitoxin System Model

Sofija Marković[1,*], Magdalena Đorđević[2], Hong-Yu Ou[3] and Marko Đorđević[1]

[1] Faculty of Biology, University of Belgrade, Belgrade, Serbia
[2] Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia
[3] School of Life Sciences & Biotechnology, Shanghai Jiaotong University, Shanghai, China
sofija.markovic@bio.bg.ac.rs

Antibiotic persistence refers to a phenomenon where a subset of genetically identical bacteria enters a dormant state, becoming highly resistant to environmental stresses. This phenomenon is crucial in understanding why biofilms, communities of bacteria attached to surfaces, often resist antibiotic treatments, leading to persistent and recurrent infections. Despite being recognized for almost a century, the precise processes triggering persister formation remain elusive.

Among the various biological systems implicated in persister formation, toxin-antitoxin systems within bacteria stand out. These systems consist of a toxic protein and its corresponding antitoxin, which neutralizes the toxin's effects. In this study, we propose a biophysical model focusing on a type I toxin-antitoxin system where the antitoxin is a small RNA molecule. Our analysis involves both theoretical calculations and computer simulations to explore the stability of the model and its behavior under deterministic and stochastic conditions.

Our model successfully reproduces two distinct states within bacterial populations: a low-toxin state associated with normal growth and a high-toxin state leading to persister formation. We analytically derive a system stability diagram, allowing us to map under which conditions the low and high toxin states coexist in an isogenic bacterial population. Furthermore, we observe a stochastic transition from low to high-toxin. This bistability in our model arises from feedback loops governing toxin production. Specifically, a positive feedback loop controls toxin dilution rate, while a negative feedback loop slows down antitoxin degradation.

Our findings have significant implications for understanding bacterial persistence mechanisms. We have shown that type I toxin-antitoxin systems may play a role in stress-induced persister formation. However, they are unlikely to account for "spontaneous" persister formation, as toxin expression is markedly reduced during normal growth phases. These insights could lead to developing new therapeutic approaches that target the specific mechanisms of stress-induced persister formation, thereby improving the effectiveness of antibiotic treatments.

Keywords: Antibiotic persistence, Toxin-antitoxin systems, Bistability, Bifurcations, Stochastic simulations.

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

Novel multi-omics deconfounding variational autoencoders can obtain meaningful disease subtyping

Zuqi Li[1], Sonja Katz[2,3,4,*] and Gennady V. Roshchupkin[2,5]

[1] Laboratory for Systems Medicine, Department of Human Genetics, KU Leuven, Leuven, Belgium
[2] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, The Netherlands
[3] Laboratory of Systems and Synthetic Biology, Wageningen University & Research, Wageningen, The Netherlands
[4] LifeGlimmer GmbH, Berlin, Germany
[5] Department of Epidemiology, Erasmus MC, Rotterdam, The Netherlands
sonja.katz@wur.nl

Unsupervised learning, particularly clustering, is crucial for disease subtyping and patient stratification. With the availability of large-scale multi-omics data, clustering algorithms can be empowered by deep learning models, e.g. variational autoencoder (VAE), to exploit the between-individual heterogeneity. However, the impact of confounders—external factors unrelated to the condition, e.g. batch effect and age—on clustering is often overlooked, introducing bias and spurious biological conclusions.

We proposed four VAE-based deconfounding approaches utilizing multi-omics data: i) removal of latent features correlated with confounders ii) a conditional variational autoencoder (cXVAE), iii) adversarial training, and iv) adding a regularization term to the loss function. Based on real-life multi-omics data from The Cancer Genome Atlas, we simulated various confounding effects (linear, non-linear, categorical, combination) and evaluated each model's performance across 50 repetitions based on reconstruction error, clustering stability, and deconfounding effectiveness, measured via the adjusted rand index (ARI).

We demonstrated a substantial impact of the artificially introduced confounder effect on patient clustering (ARI: 0.33±0.12), yet various proposed models effectively mitigated this effect, with cXVAE clearly outperforming other frameworks (ARI: 0.66±0.07). cXVAE not only accurately recovers true patient labels but also reveals meaningful pathological associations among cancer types, reinforcing deconfounded representation validity. Conversely, our study proved that some of the proposed strategies, such as adversarial training, are incapable of sufficiently removing confounders.

Our study contributes to delivering accurate patient subgrouping by not only (i) proposing novel frameworks for simultaneous multi-omics data integration, dimensionality reduction, and deconfounding of clustering, but also by (ii) benchmarking respective frameworks on open-access data to aid fellow researchers in selecting an appropriate framework that they can readily apply in a health-related settings.

Keywords: deep learning, autoencoder, multi-omics, confounders, clustering

Acknowledgement: We want to thank the supporters of this study, namely the European Union's Horizon 2020 Marie Skłodowska-Curie grant agreement (860895). Also, we would like to extend our gratitude to the co-authors making this work possible: Edoardo Saccenti (Wageningen University & Research; The Netherlands), David W. Fardo (University of Kentucky; The United States), Peter Claes (KU Leuven, University Hospitals Leuven; Belgium), Vitor A.P. Martins dos Santos (Wageningen University & Research; The Netherlands, LifeGlimmer GmbH; Germany), and Kristel Van Steen (KU Leuven, University of Liege; Belgium, University of Kentucky; The United States).

Advancing Supervised Machine Learning for scRNA-seq Data Analysis

Xin Lin[1,*], Minjie Lyu[1], Tian-yi Qiu[2], Guanglan Zhang[3], Sen Lin[1], Lou Chitkushev[3] and Vladimir Brusic[1,3]

[1] Smart Medicine Laboratory, University of Nottingham, Ningbo, China
[2] Institute of Clinical Science, Zhongshan Hospital; Fudan University, Shanghai, China
[3] Metropolitan College, Boston University, Boston, USA
xin.lin@nottingham.edu.cn
vladimir.brusic@nottingham.edu.cn

The exponential growth of single-cell transcriptomic data presents a significant challenge for the analysis of single-cell transcriptomic data. Current best practices rely on unsupervised clustering. The applications of supervised machine learning (ML) for the analysis of single-cell transcriptomic (scRNA-seq) data have increased in recent years. The main advantages of supervised ML are higher classification accuracy, and reproducibility and reliability of results, compared to unsupervised clustering. However, single-cell transcriptomic technologies are evolving rapidly, resulting in limited reproducibility of results due to changes in biological sample processing and technical differences between subsequent experimental measurements. A lack of high-quality standardized reference datasets increases the risk of model overfitting and reduces model generalization properties. Benchmarking supervised machine learning algorithms is challenging because of the lack of reference datasets.
For the advancement of scRNA-seq applications, we need high-quality annotated standardized datasets. To address the need for the deployment of supervised ML in this field, we developed a single-cell transcriptomic database of reference datasets for healthy human peripheral blood mononuclear cells (PBMC). We collected over two million single-cell data from multiple public data sources and applied advanced cell annotation methods to create multi-annotation labels in the healthy PBMC reference dataset. Each cell has labels designating cell type and subtypes, cell cycle, and cell state, along with assigned degree of belief. The annotations are based on the multi-dimensional cell ontology that we have designed. scRNA-seq data in our database were converted into a standardized format using a defined protocol that enables the direct use of data for supervised ML tasks. The data standardization pipeline and cell annotation tools are deployed within the database. The database is deployed as a publicly accessible web server for the study of single-cell PBMC.

Keywords: supervised machine learning, single cell transcriptome, transcriptome database, cell annotation.

Acknowledgements: This work was supported by the University of Nottingham Ningbo China High-Flyer Scholarship, code: 2106HFB.

Development of automated pharmacogenetic report for evaluating possible side effects of acute lymphoblastic leukemia therapy

Suvorova Y.[1,*], Monakhova A.[1], Gurzhikhanova M.[2], Zaigrin I.[1] , Antonov I.[1], Musharova O.[1], Klimuk E.[1] and Severinov K.[1]

[1] Limited Liability Company »Biotechnology campus«, Moscow, Russia
[2] Dmitry Rogachev Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia
ysuvorova@biotc.ru

Pharmacogenetics is the study of the genetic basis of individual responses to drugs. Adverse drug reactions (ADR) effects occur when an individual with a certain risk genotype gets a certain drug. The majority of risk genotype carriers are unaware of their genetic predisposition. Information about an association between genotypes and drug responses helps drug prescription and dosage. It is particularly important when patients receive many drugs during the treatment courses, such as treatment from acute lymphoblastic leukemia (ALL) in Dmitry Rogachev Center of Pediatric Hematology, Oncology and Immunology. The goal of our work was to develop a personal pharmacogenetic report for childhood ALL patients based on WGS data.

In collaboration with the doctors we have compiled a list of drugs and the most relevant side effects of the treatment. Based on literature and public resources analysis we collected a database of genomic variants and haplotypes associated with ADR of ALL therapy. Our pharmacogenetic report lists risks of side effects of 25 drugs based on haplotypes of 13 pharmacogenes and almost 100 short variants.

To define haplotypes of the pharmacogenes we have developed a bioinformatics pipeline that includes public tools and our own code. The pipeline was validated on reference samples from the GetRM dataset and 1000 WGS samples from the "100,000+Me" Initiative. A Django-based system was developed for the ADR risks calculation and reports generation. The system takes WGS data as an input, creates predictions for each drug and generates a final pdf report for the patient. Genotypes, haplotypes and predictions are stored in a database for further analysis. Over 50 reports were already issued to the chemotherapy specialists to assist them during the drug and dosage selection process.

Keywords: pharmacogenetics, acute lymphoblastic leukemia, WGS, personalized medicine

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

The danger of powerful mitochondria: life-history traits shape the evolution of bird mtDNA

Gusarov Y.S.[1,*], Burskaya V.O[2], Bushuev A.V.[3], Mikhailova A.G.[1], Efimenko B.E.[1], Gunbin K.V[1] and Popadin K.Y.[1,4]

[1] Immanuel Kant Baltic Federal University, Kaliningrad, Russian Federation
[2] University of Antwerp, Antwerp, Belgium
[3] Lomonosov Moscow State University
[4] Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland
yurgusguss@mail.ru

A>G mutation in the mitochondria heavy chain is one of the most discussed types of mutational signatures: previous studies prove its association with mtDNA chemical damage, caused by elevated energy production level. Here we study A>G mutation patterns in birds - creatures with highest energy consumption in the world. We demonstrate a highly expected excess of A>G mutation frequency in birds compared to mammals. In order to understand biochemical mechanisms, which govern A>G mutations in birds, we describe a list of associations between A>G mutation frequencies and life history traits of birds.First, we show that unlike mammals, birds demonstrate no connection between frequency of A>G mutations and most obvious chemical damage correlates: body mass, lifespan or basal metabolic rate (BMR). It is a surprising result, which we interpret as a sign of a highly optimized electron-transport chain, which produces a very stable level of chemical damage agents in a wide range of "regular" conditions. However, according to our results, there is a list of "irregular" conditions, which cause significant change of A>G mutation frequencies. Decrease of A>G mutation is caused by loss of flight: this change is in line with drastic decrease of energy consumption. We have found an increase of A>G mutation rate in diving birds (diving hypoxia is known to produce chemical damage) and in long range migrators (which are known for outstanding peak metabolism levels). The results are significant after phylogenetic generalized least squares correction and show strong phylogenetic inertia.

Keywords: mitochondria, mtDNA, birds, evolution, mutagenesis

Acknowledgements: The study is supported by RSF grant (No. 21-75-20143).

A systematic characterization of genotype-to-phenotype relationships in mouse and human

Yury Barbitoff*, Nadezhda Pavlova, Polina Bogaichuk and Alexander Predeus

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

Reconstruction of the genotype-to-phenotype relationship network is one of the major goals of modern genomics. Development of ontologies for accurate and formal phenotype description enables a systematic comparison of these relationships between species. In this work, we employed data on genotype-to-phenotype associations from the Human Phenotype Ontology (HPO) and Mouse Genome Informatics (MGI), as well as genome-wide associations from the UK Biobank (UKB) cohort to investigate the similarities and dissimilarities in the architecture of genotype-to-phenotype associations in mice and humans.

Comparison of the sets of upper-level Mammalian Phenotype Ontology (MP) terms for 16,985 orthologous gene pairs showed that only 24.6% (4,184) of all gene pairs were annotated with phenotype terms in both species. Even more surprisingly, only a handful of 15 gene pairs were annotated with exactly identical terms in human and mouse. Of the remaining ones, as many as 385 genes were associated with non-overlapping sets of phenotypes in humans and mice, and as many as 3,784 genes had both concordant and discordant gene-trait associations.

Despite such large differences in genotype-to-phenotype relationships for individual genes, we found that the overall architecture of the genotype-to-phenotype network was similar for the two species. For instance, the pleiotropic effects of orthologous genes were well correlated (Spearman's $p = 0.28$) when using upper-level phenotype term count as a measure of pleiotropy. In both species, higher degree of pleiotropy also correlated with various gene-level properties, such as greater evolutionary constraint of a gene and its broader expression across tissues. Similar trends were observed when using UKB genome-wide associations to estimate the degree of pleiotropy for human genes.

Taken together, our observations highlight important discrepancies between phenotype description in humans and mice. At the same time, our analysis suggests that the organization of genotype-to-phenotype networks is largely similar for highly related species.

Keywords: genotype-to-phenotype; complex traits; phenotype ontology; HPO; MPO

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

INMTD: integrative clustering with 2D genotypes and 3D facial images in the presence of confounders

Zuqi Li[1,2,*], Sam F. L. Windels, Seth M. Weinberg, Mary L. Marazita, Susan Walsh, Mark D. Shriver, Noël Malod-Dognin, David W. Fardo, Peter Claes, Nataša Pržulj and Kristel Van Steen[1,3]

[1] Department of Human Genetics, KU Leuven, Leuven, Belgium
[2] Medical Imaging Research Center, UZ Leuven, Leuven, Belgium
[3] GIGA-R Medical Genomics, University of Liège, Liège, Belgium
zuqi.li@kuleuven.be

By integrating genomic and facial images, we can achieve a more comprehensive, multi-view clustering of individuals. Among the various approaches for multi-view clustering, integrative nonnegative matrix tri-factorization (NMTF) has emerged as advantageous in learning low-rank embedding of samples and features and interpreting these representations. Incorporating 3D imaging is challenging, but here is where nonnegative Tucker decomposition (NTD) can come in. In this work, we introduce a novel multi-view clustering method based on both NMTF and NTD, namely INMTD, that integrates genotypes and 3D facial images to generate unconfounded subgroups of individuals. Indeed, there is a need to handle unwanted drivers of clusterings (i.e. confounders). We applied our method to real-life multi-view data on 4680 individuals from a US cohort. Several confounders were also available, such as age, sex and height. When removing these factors, one would expect population structure to be the prevailing driver for the heterogeneity. In particular, INMTD generates three embedding matrices for 1) samples, 2) SNPs and 3) facial landmarks. The biological relevance of these embeddings was investigated in several ways. For 1), most sample embedding vectors were statistically significantly associated with ancestry axes or confounders. By removing confounded vectors in the sample embedding, we derived an unconfounded clustering with better internal quality and stronger association with population structure; the genetic and facial annotations of each derived subgroup highlighted different physiological or morphological characteristics. Regarding 2), clusters of embedded SNPs showed good enrichment of genes and Gene Ontology terms. For 3), the segmentation on the facial embedding improved cophenetic correlation compared to earlier reports on the same data. Projecting SNPs and facial landmarks to the sample embedding space revealed known and novel SNP-face biological relationships. In conclusion, INMTD can effectively integrate omics data and 3D images for unconfounded clustering with biologically meaningful interpretation.

Keywords: multi-view clustering, matrix factorization, confounder, population genetics, facial images

Acknowledgement: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreements No 813533 (MLFPM) and No 860895 (TranSYS).

Transgenerational transmission of post-zygotic mutations suggests symmetric contribution of first two blastomeres to human germline

Yeongjun Jang[1], Livia Tomasini[2], Taejeong Bae[1],
Anna Szekely[3], Flora M. Vaccarino[2,4,5,*] and Alexej Abyzov

[1] Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
[2] Child Study Center, Yale University, New Haven, CT 06520, USA
[3] Department of Neurology, Yale University, New Haven, CT 06520, USA
[4] Department of Neuroscience, Yale University, New Haven, CT 06520, USA
[5] Yale Kavli Institute for Neuroscience, New Haven, CT 06520, USA
abyzov.alexej@mayo.edu
flora.vaccarino@yale.edu

Little is known about the origin of germ cells in humans. We previously leveraged post-zygotic mutations to reconstruct zygote-rooted cell lineage ancestry trees in a phenotypically normal woman, termed NC0. Here, by sequencing the genome of her children and their father, we analyzed the transmission of early pre-gastrulation lineages and corresponding mutations across human generations. We found that the germline in NC0 is polyclonal and is founded by at least two cells likely descending from the two blastomeres arising from the first zygotic cleavage. Analyses of public data from several multi-children families and from 1,934 familial quads confirmed this finding in larger cohorts, revealing that known imbalances of up to 90:10 in early lineages allocation in somatic tissues are not reflected in transmission to offspring, establishing a fundamental difference in lineage allocation between the soma and the germline. Analyses of all the data consistently suggest that germline has a balanced 50:50 lineage allocation from the first two blastomeres.

Keywords: somatic mosaicism, mutations, cell lineage tracing, development, germline

Acknowledgement: We are grateful to member of NC0 family that participated in this study by donating tissue and/or blood samples. We are grateful to members of families in SSC that donated blood samples for WGS and to Simons Foundation that provided access to the data (approved project 2343.4). This work was funded by the NIH Common Fund SMaHT program (grants UG3 NS132128 and UG3 NS132146) and by the Simons Foundation (grant 399558). Y.J. was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (grant number 2022R1A6A3A03055692).

Advances in a comprehensive understanding of alpha-1 antitrypsin glycosylation-data mining within recent publications

Anđelo Beletić[1,*], Irena Trbojević-Akmačić[1], Jasminka Krištić[1] and Gordan Lauc[1,2]

[1] Genos Ltd, Glycoscience Research Laboratory, Zagreb, Croatia
[2] Faculty of Pharmacy and Biochemistry, University of Zagreb, Zagreb, Croatia
abeletic@genos.hr

Alpha-1 antitrypsin (AAT) is the serin protease archetype, with biological roles shedding from the neutrophil-derived protease activities control to a comprehensive innate immunity modulation via interactions with the numerous proteins, cytokines, and cells. UniProt Knowledgebase® reports three glycosylation sites for AAT, with attached N-glycans contributing to approximately 12% of the total molecular mass and having a crucial role in immunomodulatory activities. The PubMed database was searched for "alpha-1 antitrypsin" and "glycosylation" using the following filters: full text available, the publication date of 10 years, and the preprints excluded. The search retrieved 73 results, and after the content analysis by the authors, 60 remained relevant, six reviews and 54 original research articles. Most studies used human samples (24 used serum/plasma, 23 cell cultures, four cerebrospinal or other non-standard sample fluids, and three studies combined the different sample types). There were also two studies on animals and one on plant models. Analysis of AAT glycoforms was the core of lab methodology in 52 studies, while six studies analyzed the released glycans. Thirteen studies focused on analytical protocols' development or optimization. The glycosylation-related AAT structural and functional properties were among the aims of 25 studies, while 10 papers assessed these features from a glycoengineering perspective. The pathophysiological mechanisms relating to AAT glycosylation features were studied in 10 papers, and 20 studies identified AAT glycoforms as biomarker candidates, mainly in the oncology field (pancreatic, liver, oral, and ovarian cancer). These pioneer data mining results indicate the availability of comprehensive data about AAT glycosylation, thereby establishing a solid basis for further scientific and innovation efforts.

Keywords: data mining, alpha-1 antitrypsin, glycosylation

WGS approach to identify potential genetic modifiers in Glycogen Storage Disease Ib

Skakic A[1,*], Parezanovic M[1], Pavlovic Dj[1], Stevanovic N[1], Andjelkovic M[1], Klaassen K[1], Ugrin M[1], Komazec J[1], Spasovski V[1], Djordjevic M[2], Pavlovic S[1] and Stojiljkovic M[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia;
[2] Mother and Child Healthcare Institute „dr Vukan Cupic", School of Medicine, University of Belgrade, Serbia
anita.skakic@imgge.bg.ac.rs

Glycogen Storage Disease Ib (GSD Ib) is a rare metabolic disorder characterized by a deficiency of glucose-6-phosphate translocase, leading to metabolic disruptions and neutropenia. Varying severity and progression of neutropenia were detected among individuals with the same genotype, indicating a complex genotype-phenotype correlation. We aim to explore potential modifier genes influencing neutropenia in GSD Ib, focusing on five patients with the homozygous c.1042_1043delCT variant in the *SLC37A4* gene. These patients exhibit diverse neutropenia profiles, with two displaying mild and intermittent neutropenia, while the remaining three develop severe and persistent neutropenia.

Whole genome sequencing (MGISEQ-G400, BGITech) was conducted on five unrelated subjects, all presenting with previously identified pathogenic homozygous *SLC37A4* variant. We followed GATK best practices for genomic data processing to identify genetic variations associated with observed clinical differences. A unique pipeline was constructed focusing on neutropenia-related genes and genes involved in glucose-6-phosphate metabolism, neutrophil function pathways, immune system regulation, ER stress, and UPR response.

In patients with severe neutropenia, we identified two heterozygous variants (c.-70G>C and c.96T>C, p.Thr32Thr) in the *JAGN1* gene, which is essential for neutrophil differentiation. Additionally, severe neutropenia patients had variants in *CTLA4* (c.49A>G, p.Thr17Ala) and *TGFB1* (c.29C>T, p.Pro10Leu), genes involved in immune regulation and cell survival and which have previously been recognized as modifier genes in various immunological conditions.

This research underscores the potential significance of modifier genes in shaping the diverse course of neutropenia in GSD Ib, highlighting the need for further functional studies to elucidate the precise roles of these variants in disease presentation. Investigating potential genetic modifiers can provide valuable insights into the molecular base of the disease and guide future research focused on developing customized therapeutic approaches for the specific neutropenic phenotype.

Keywords: GSD Ib, WGS, modifier genes

Acknowledgment: This research was supported by the Science Fund of the Republic of Serbia, #Grant No. 6999, New concept for treatment of glycogen storage disease Ib and diabetes mellitus type 2: small molecule compounds able to adjust glucose level through binding glucose-6-phospate translocase - GlucoAdjust.

Towards Head and Neck Myeloid Cells Atlas

Dragana Dudic[1,*], Diana Domanska[2], Nicolina Sciarffa[3], Francisca Hofman-Vega[4], Serafina Reif[5] and Nico Trummer[5]

[1] Faculty of Computer Science and Informatics, University Union Nikola Tesla, Belgrade, Serbia
[2] Department of Pathology, University Hospital, Oslo, Norway
[3] Advanced Data Analysis Group, Ri.MED Foundation, Palermo, Italy
[4] Department of Otorhinolaryngology, University Hospital Essen, Essen, Germany
[5] TUM School of Life Sciences, Technical University of Munich, Freising, Germany
ddudic@unionnikolatesla.edu.rs

Head and neck cancer is the seventh most common cancer in the world with squamous cell carcinoma as the most common histology. This heterogeneous group of tumors with aggressive malignancy is characterized by a specific tumor microenvironment in which myeloid cells dominate with its role to initiate antitumor response.
In order to reveal principles of immunity for different regions affected with head and neck cancer, we are creating the atlas of head and neck myeloid cells based on publicly available and in-house single-cell transcriptome datasets created with different single-cell transcriptome sequencing platforms. Current version of head and neck myeloid cells atlas is comprised of core atlas and extended atlas, where the core atlas includes 14 datasets with 102 patients and 582384 cells, created with 10x Genomics sequencing platform, while extended atlas includes 3 datasets with 17 patients and 131380 cells, created with other single-cell sequencing methods like BD Rhapsody, SmartSeq2 and In-Drop.
Dataset specific preparation has been done in R and Python. In order to automate the process for the rest of the analysis, we created a SIMBA pipeline which is publicly available on GitHub. We performed automatic and manual annotation of the head and neck myeloid cells atlas. Automatic annotation is conducted in two ways, using SingleR with multiple datasets and Celltypist with majority voting. Manual annotation is based on markers defined by domain experts. All obtained annotations are publicly available through CellxGene platform.
Our future work includes expanding the head and neck myeloid cell atlas with novel datasets, deconvolution based on bulk transcriptome data available in TCGA database and addition of spatial transcriptomics data.

Keywords: bioinformatics, single cell transcriptomics, scRNA-seq, head and neck cancer

Acknowledgement: This publication is based upon work from COST Action Mye-InfoBank CA20117, supported by COST (European Cooperation in Science and Technology).

Silicon Affects The Expression Of Conserved And Novel Cucumber miRNAs In Response To Copper Stres

Dragana Bosnić*, Gordana Timotijević, Dragana Nikolić and Jelena Samardžić

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Belgrade, Serbia
dragana.bosnic@imgge.bg.ac.rs

Environmental pollution with heavy metals such as copper (Cu) severely hinders optimal plant growth and development. Silicon (Si) is a plant nutrient that can improve plant health through diverse mechanisms. microRNAs (miRNAs) are recognized as effective regulators of various processes in plants, including stress responses, however, their involvement in Si-protective effect remains unrevealed. To clarify the molecular pathways involved in the protective effect of Si, small RNA-seq analyses (Illumina Sequencing SE50) was performed on cucumber plants treated with silicon (Si) alone or with a high concentration of Cu (Cu+Si). The small RNA tags were mapped to the reference sequence by Bowtie tool, following no-mismatch or one mismatch criterion, to analyze their expression and distribution. Identification of known miRNAs among the mapped small RNA tags was accomplished by miRBase22.0 database. miREvo and mirdeep2 were exploited to predict novel miRNAs by analyzing the secondary structure, Dicer cleavage site, and minimum free energy of the small RNA tags unannotated in the previous steps. The prediction of the target gene of miRNA was performed by psRobot. Differential expression analysis of two treatments was done using edgeR, with the default threshold qvalue < 0.05 and |log2(fold change)| > 1.
A total of 71 miRNAs were identified in cucumber plants. Twenty miRNAs that were significantly differentially expressed, including seven novel miRNAs were clustered to find similar expression patterns. miR398 and miR408, which are known to target Cu-proteins important for Cu metabolism and transport, were significantly downregulated in Cu+Si treatment. In contrast, miR156, miR164, mir167, miR394, and miR477 were upregulated in Cu+Si. Additionally, two novel miRNAs (Novel_15 and Novel_19) were highly expressed and specific to Cu+Si, while, the Novel_12 miRNA was specific to Si treatment. The putative target genes of the differentially expressed miRNAs were subjected to Gene Ontology (GO) enrichment analysis, which revealed localization and transport as the most significant GO terms in this study.

Keywords: silicon, copper, plants, sequencing, miRNAs

Acknowledgement: This study has been funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (grant number 451-03-66/2024-03/200042).

A comparative transcriptomic analysis of mouse DM1 models' skeletal muscles

Dušan Lazić[1,*], Vladimir M. Jovanović[2], Jelena Karanović[1], Dušanka Savić-Pavićević[1] and Bogdan Jovanović[1]

[1] University of Belgrade-Faculty of Biology, Center for Human Molecular Genetics, Belgrade, Serbia;
[2] Human Biology and Primate Evolution, Department of Biology, Chemistry and Pharmacy, Freie Universität Berlin, Berlin, Germany
dusan.lazic@bio.bg.ac.rs

Myotonic dystrophy type 1 (DM1) is a rare, incurable multisystemic disease, with the main symptoms being skeletal muscle weakness, atrophy, and myotonia. It is caused by CTG expansion in the 3' UTR of the DMPK gene whose RNA acquires toxic functions and sequesters MBNL proteins, resulting in globally altered RNA metabolism. Despite having many mouse models with different phenotypes, none of them has been able to fully recapitulate the phenotype and molecular pathogenesis of DM1. To map transcriptomic differences among various mouse DM1 models, we systematically analyzed gene expression in their skeletal muscles.

We retrieved all publicly available RNA-seq datasets from mouse models expressing expanded CTG repeats and Mbnl knockout models. Our workflow with unified parameters consisted of preprocessing, and differential gene expression analysis (DESeq2). Additionally, gene co-expression networks (WGCNA), were focused on the CTG repeat-expressing model that was most commonly used and had the largest number of biological replicates (HSALR), where network nodes were represented by a union of dysregulated genes from all analyzed datasets.

In models expressing CTG repeats, the average number of up- (787) and down-regulated (642) genes was greater compared to Mbnl knockouts (676 and 380; log2FC>1, padj>0.05). WGCNA recovered three modules of strongly co-expressed genes (adjacency>0.5). The turquoise module, sharply correlated with muscle type, was associated with extracellular space and muscle development. The midnightblue module consisted of Mup gene family members, while the brown module was associated with the immune response.

Our results revealed pathway changes in DM1 skeletal muscles, where immune pathways in muscle homeostasis and development are intriguing as molecular targets for further investigation. Furthermore, gene expression patterns separated Mbnl knockouts from models expressing CTG repeats, indicating the significance of mouse model choice for basic and preclinical research.

Keywords: Myotonic dystrophy type 1, comparative transcriptomics, mouse models, co-expression networks

Acknowledgment: This research was supported by the Science Fund of the Republic of Serbia, Grant number 7754217, READ-DM1

Analysis of RNA Secondary Structural Elements Using the RNAsselem Python Package

Fedor M. Kazanov[1], Evgenii V. Matveev[2,3,4] , Gennady V. Ponomarev[2,3], Dmitry N. Ivankov[2] and Marat D. Kazanov[2,3,4,5,*]

[1] "Foxford" High School, Moscow, Russia
[2] Skolkovo Institute of Science and Technology, Moscow, Russia
[3] A.A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia
[4] Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia
[5] Sabanci University, Istanbul, Turkey
mkazanov@gmail.com

RNA, a molecule essential for numerous biological functions including genetic information storage, gene regulation, and catalytic activity. Its ability to form specific secondary and tertiary structures is crucial for its functionality. The field of RNA secondary structure analysis is advancing rapidly, owing to enhancements in both experimental and computational techniques. Recent research, especially focusing on RNA viruses like HIV, influenza, and SARS-CoV-2, demonstrates how viral RNA structures critically influence stages of the viral life cycle such as replication and immune evasion. This is exemplified by structures like the HIV-1 Rev response element and the SARS-CoV-2 frameshift stimulation element, which are pivotal in functional interactions essential for viral replication and protein synthesis.
However, existing RNA analysis bioinformatics tools have limitations, mostly focusing solely on nucleotide pairing and lacking the recognition of common patterns like hairpins, bulges, and pseudoknots. Here, we present a Python package specifically designed for analyzing RNA secondary structural elements in viral genomes. This tool facilitates the identification of common secondary structure patterns such as hairpin loops, internal loops, and pseudoknots, among others, and provides a framework for analysis of these elements to get insights into their properties. RNAsselem Python package is available at: http://github.com/KazanovLab/RNAsselem.

Keywords: RNA secondary structure, RNA viruses, hairpin, stem, loop.

Acknowledgement: This research was funded by Russian Science Foundation, grant number 22-14-00132.

Detection, identification and quantification of target DNA sequence in soybean event GTS 40-3-2

Ilma Mujković[1,*], Kasim Bajrović[2] and Teodora Andrejić[3]

[1] Faculty of Science, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
[2] Institute for genetic engineering and biotechnology, University of Sarajevo, Sarajevo, Bosnia and Herzegovina
[3] Analysis, Novi Beograd, Serbia
ilma.mujkovic@gmail.com

The soybean event OECD line MON-Ø4Ø32-6, also known as GTS-40-3-2 or Roundup Ready® soybean (RRS), was developed to tolerate glyphosate, the active ingredient in the herbicide Roundup®. This genetically modified (GM) soybean contains a sequence from the *Petunia hybrida* chloroplast coding transit peptide and the *Agrobacterium tumefaciens* 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS) gene. In Bosnia and Herzegovina, no GMO has been approved for cultivation or use as food, but GTS 40-3-2 is one of eight types of genetically modified soy approved for use as animal feed. The aim of this study was the identification and quantification of potentially genetically modified soybeans marketed in Bosnia and Herzegovina using polymerase chain reaction (PCR). DNA extraction from ten soybean samples was performed using the GENE*Spin* kit (Eurofins GeneScan GmbH) according to the manufacturer's protocol in compliance with EN ISO/IEC 17025:2017 standards. To determine the presence of GMOs for two *screening* elements, P-35S and t-NOS, *end-point* PCR (GMO*Screen* 35S/NOS kit) was applied, and for the third element, FMV (P-34S), *Real-Time* PCR (GMO*Screen* RT35S/NOS/FMV-IPC kit) was used, while the amplified sequences were analysed in the manufacturer's analytical program. Quality assurance of the test results was ensured with positive control DNA, specifically plasmid DNA pGSE211 (RRS). The *end-point* PCR method for detecting the P-35S promoter and t-NOS terminator in this study showed identical results to the *Real-Time* PCR *screening* method using kits from the same manufacturer. After detecting the P-35S promoter and t-NOS terminator, *Real-Time* PCR confirmed that it was GM soybean GTS-40-3-2. Soy sample 4S contained 2.27% RRS, and sample 7S contained 1.07%. Since these values exceed the threshold of 0.9%, according to the GMO Law of Bosnia and Herzegovina, they require labeling to indicate genetically modified material.

Keywords: *Roundup Ready* soybean, GMO, *Real-Time* PCR, quantification

Fine-tuning RNA-seq alignment parameters for *Danio rerio* genome

Jelena Kusic-Tisma*, Mila Ljujić, Bojan Ilić and Aleksandra Divac Rankov

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Belgrade, Serbia
jkusic@imgge.bg.ac.rs

In RNA-seq analysis, mapping raw reads to the reference genome using splice-aware aligners is a matter of personal preference. However, the distribution of reads among different mapping metrics strongly depends on appropriate parameters selection.
In this study we conducted a thorough analysis of parameter impact on alignment metrics using STAR aligner in 12 samples of PE150 RNA-seq data from zebrafish. When setting parameters we considered factors such as insert size distribution of pre-processed library data by fastp and the unique features of the reference genome (e.g., gene density and intron size distribution).
Average number of input reads per sample were 25097912. Adjusting parameters led to significant improvements in mapping metrics, notably an increase in the percentage of *uniquely mapped reads* from 86.08% to 94.19%. A minor rise was observed in the percentage of reads mapped to multiple loci, with figures of 4.72% and 3.62% respectively. The quantification of reads by genomic origin has revealed an increase in uniquely mapped reads allocated to exonic regions , on average by 2,034,563 reads per sample.
In addition, we've compared default and adjusted metrics of multi-sample 2-pass mapping, which are important when analysing differential transcript usage. In the 2-pass mode, we employed 1st pass junction files that were purged of probable false positives, such as junctions within the mitochondrial genome or those crossed by multi-mapping reads. Qualimap-rna analysis of 2-pass mapping revealed an increase in the usage of novel splice sites, as expected. However, default parameters showed higher rates of multimapping, noFeature, or ambiguous reads compared to adjusted parameters.
Running alignment with default options generally performs well initially. However, proper parameter selection tailored to the characteristics of the raw library data and reference genome significantly improves alignment metrics, especially for model organisms like zebrafish.

Keywords: RNA-seq analysis, STAR aligner, mapping parameters, zebrafish

Acknowledgement: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-66/2024 03/200042)

Exploring Genetic Variant Selection Algorithms for Enhanced Genotyping Assays in Personalized Medicine

Katarina Kruščić[1], Ivan Životić[2], Maja Živković[2] and Tamara Đurić[2]

[1] Faculty of Biology, University of Belgrade, Belgrade, Serbia
[2] Institute of Nuclear Sciences "Vinča", National Institute of the Republic of Serbia, Laboratory for Radiobiology and Molecular Genetics, University of Belgrade, Belgrade, Serbia
katarina.kruscic@bio.bg.ac.rs

The integration of complex genetic information into healthcare systems is vital for advancing personalized medicine. Linkage disequilibrium (LD) facilitates obtaining significant information with fewer genetic variants designated as tag variants. Accurate tag variant selection enhances assay efficiency by maximizing information yield with fewer markers. Previous algorithms based on haplotype blocks prompted the development of newer algorithms like BigLD, offering improved LD estimation. However, better recombination estimation methods are needed due to incomplete alignment with recombination sites. Expanding search regions by ±100kb around gene boundaries reveals regulatory elements that impact gene expression. This study aims to compare LD-Select and LmTag algorithms, focusing on imputation coverage and functional aspects. Research was conducted on the two genes (*Gclm* and *Nqo1*) from the Nrf2/ARE signaling pathway, important in inflammatory conditions. LmTag, considering minor allele frequency (MAF), LD strength (r^2), and genomic distance, aims to enhance performance and functional relevance compared to LD-Select. Results indicate that while increasing bin width enhances cumulative scores, it reduces imputation coverage. Optimal bin widths were determined for specific genes: k=50 for *Gclm* gene, and k=10 for *Nqo1* gene. Extending gene regions by 100kb improves imputation coverage using both algorithms. For instance, for the *Gclm* gene and its promoters, LD-Select suggests 5 tag variants covering 97.30% of the region. Expanding the observation region reveals 6 tag variants with 100% coverage, while LmTag selects 5 variants covering 94.59%. For the *Nqo1* gene, LD-Select suggests 3 tag variants covering 95% of the region, with LmTag offering 3 variants with similar coverage. This study underscores the importance of selecting representative variants and offers insights into algorithmic development, with potential implications for personalized medicine in diverse phenotypic conditions. Further experimental validation is essential to corroborate these findings and enhance their clinical utility.

Keywords: tag variants, linkage disequilibrium, LD-Select, LmTag, BigLD

Acknowledgement: This research was supported by the Science Fund of the Republic of Serbia, grant number #Grant no. 7753406, Identification and functional characterization of extracellular and intracellular genetic regulators of ferroptosis related processes in multiple sclerosis – FerroReg.

Bioinformatics insights into genes encoding heat-resistant obscure (Hero) proteins and their role in cardiovascular diseases: a regulatory SNP analysis

Vladislav Shilenok[1], Anna Dorofeeva[1,*], Irina Shilenok[1], Ksenia Kobzeva[1] and Olga Bushueva[1,2]

[1] Laboratory of Genomic Research, Research Institute for Genetic and Molecular Epidemiology, Kursk State Medical University, Kursk 305041, Russia
[2] Department of Biology, Medical Genetics and Ecology, Kursk State Medical University, Kursk 305041, Russia
annadorofeeva1809@gmail.com

Cardiovascular diseases (CVDs) represent a significant global health burden, claiming over 20 million lives annually according to the World Heart Federation. While considerable progress has been made in understanding the molecular pathology of CVDs, significant knowledge gaps persist. Given the pivotal role of chaperones in cellular function and disease, there is growing interest in exploring their involvement in CVDs. We aimed to conduct a bioinformatic analysis to investigate the cardiovascular contributions of genes encoding heat-resistant obscure (Hero) proteins with chaperone activity.

The study included genes encoding Hero: *C9orf16*, *C11orf58*, *SERBP1*, *SERF2* and *C19orf53*. Various bioinformatic resources, including SNPinfo Web Server, The Cerebrovascular Disease Knowledge Portal (CDKP), The Cardiovascular Disease Knowledge Portal (CVDKP), HaploReg, rSNPBase, RegulomeDB, atSNP, Gene Ontology, QTLbase, Blood eQTL browser, were employed to annotate the cardiovascular relevance of Hero proteins and their genes.

All analyzed genes, except for *BEX3*, exhibit high regulatory potential of tagSNPs. Hero genes are expressed in cardiovascular and neural tissues, with their expression potentially modulated by cis-eQTL effects. Interestingly, certain tagSNPs also influence the expression of genes associated with CVDs. Additionally, we observed correlations between Hero gene tagSNPs and methylation levels of CpG sites, particularly in blood and brain cells, via cis-mQTL effects. Moreover, these tagSNPs significantly influence binding with transcription factors involved in key biological processes implicated in CVD pathogenesis, including vasculogenesis, response to oxidative stress and hypoxia, inflammation regulation, cytokine production, apoptosis, neurogenesis, and neurodifferentiation. Our bioinformatic analysis suggests that genes *C9orf16*, *C11orf58*, *SERBP1*, *SERF2*, and *C19orf53* play roles in various subtypes of ischemic stroke and the severity of stroke, intracranial hemorrhage, coronary artery disease/myocardial infarction, peripheral artery disease, and hypertension. Additionally, we identified various intermediate phenotypes influenced by these tagSNPs, such as lipoprotein levels, dyslipidemia, body mass index, and atrial fibrillation.

Our findings provide compelling *in silico* evidence of Hero-proteins' involvement in CVDs and their intermediate phenotypes, highlighting their potential significance in heart and vessel disease pathogenesis.

Keywords: Hero (heat-resistant obscure) genes; SNP; functional annotation; cardiovascular diseases

Acknowledgement: This research was funded by the Russian Science Foundation (22–15–00288).

Pharmacogenetics-based Dosing Algorithm for Acenocoumarol in the Serbian Population

Rakicevic Ljiljana[1,*], Kovac Mirjana[2,3] and Radojkovic Dragica[1]

[1] Institute of molecular genetics and genetic engineering, University of Belgrade, Belgrade, Serbia
[2] Faculty of Medicine, University of Belgrade, Belgrade, Serbia
[3] Blood transfusion institute of Serbia, Hemostasis department, Belgrade, Serbia
lili@imgge.ac.bg.rs

Pharmacogenetics, as a discipline which correlates genetics of an individual and the effects of drugs, has given new possibilities for personalized approaches in medicine. It is possible to design algorithms to predict the effects of a certain therapeutic by analysing relevant genetic variants as well as non-genetic factors which may influence therapy. It has been shown that algorithms designed in this way allow for better prediction in comparison to traditional trial and error method and represent a more cost-effective approach for health systems. Additionally, the contribution of factors affecting therapy may vary markedly between different ethnic groups. One of the most considered drugs in pharmacogenetics are coumarins (warfarin, acenocoumarol, phenprocoumon), anticoagulation drugs, used in treating and preventing thromboses.

This work was aimed to design pharmacogenetics-based algorithm for acenocoumarol. Assuming that population-specific algorithm may take advantage over models used in a generalized manner, we aimed to design a mathematical model for predicting individual drug dosage in the Serbian population based on clinical-demographic and genetic data.

Patients with stable acenocoumarol maintenance dose (N = 200) were divided into two cohort – derivation cohort (N = 100) and testing cohort (N = 100) – on a random basis. On the derivation cohort multiple regression analysis was applied in order to select predictors to be used for estimating the individual dose of acenocoumarol and to derive a model for dose prediction. The testing cohort was used for assessing the quality of the derived model.

Mathematical model for predicting individual acenocoumarol dose was designed and its unadjusted R2 was 61.8. In addition to genetic factors (*VKORC1*2*, *CYP2C9*2*, *CYP2C9*3*), we identified age, weight and gender of the patients as significant predictors of drug dosage. In comparison with the model given by other authors our model showed better prediction of individual acenocoumarol dose for patients in the Serbian population.

Keywords: pharmacogenetics, algorithms, personalized medicine

Acknowledgement: This work was supported by the Ministry of science, technological development and innovation of the Republic of Serbia (contract No: 451-03-66/2024-03/200042).

A Transcriptomic Meta-analysis of Carbon Nanomaterials Toxicity on Lung Tissue

Mariana Seke, Ivan Jovanović, Nataša Mačak*,
Maja Živković and Aleksandra Stanković

Institute for nuclear sciences ''Vinča'', National institute of the Republic of Serbia, University of Belgrade, Serbia
natasa.macak@vin.bg.ac.rs

Nowadays, carbon nanomaterials (CNMs) are produced on an industrial scale and exposure to them can cause serious health issues. In this study, the long-term effects on the lung tissue transcriptome after administration of five CNMs: graphene oxide, reduced graphene oxide, small multi-walled carbon nanotubes (MWCNT), large MWCNT, and carbon black were examined. Since these CNMs differ physicochemically, the aim was to investigate whether these CNMs induced similar or different transcriptomic signatures.
Publicly available transcriptomic datasets from the Gene Expression Omnibus database, under accession numbers GSE159707, GSE35284, and GSE55286 were used to extract treatment data. Two doses of CNMs were chosen, 0 μg/mouse in the control group and 18 μg/mouse in the treatment group. Transcriptomic profiling was performed four weeks after each treatment. The analysis of data was done using GEO2R tool, which employs Linear Models for Microarray Analysis (limma) R package to identify differentially expressed genes (DEGs) between the control group and each CNM-treated group. The DEGs were identified with a cut-off nominal p value of 0.05. Subsequently, a meta-analysis of the identified DEGs was performed using iPathwayGuide tool (Advaita Bio) to explore if the enriched signalling pathways or enriched diseases, are common for all five CNM treatments.
The results have shown an enrichment of biological processes associated with immunological response in all CNMs treatments. Additionally, the research demonstrated that all CNMs treatments commonly induced Il-17 signalling pathway. Moreover, two disorders were identified as statistically significant: alpha-1-antitrypsin deficiency and chronic obstructive pulmonary disease. These findings are consistent with histopathological studies in mice.
In conclusion, physicochemically distinct CNMs could have similar adverse biological effects. Furthermore, the findings have shown that a single CNM treatment could have long-term impacts. This suggest the potential of a targeted profiling of CNM exposed subjects to predict the detrimental outcome and timely employ prophylactic strategies.

Keywords: carbon nanomaterials, lungs, transcriptome, bioinformatics

Acknowledgement: This research was supported by the Serbian Ministry of Science, Technological Development, and Innovation, Grant No. 451-03-66/2024-03/ 200017

Aggregation of LEA proteins from *Ramonda serbica*: *in silico* vs. *in vitro*

Ana Pantelić[1], Tatiana Ilina[1], Dejana Milić[1],
Milan Senćanski[2] and Marija Vidović[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
[2] Institute of Nuclear Sciences VINCA, National Institute of Serbia, University of Belgrade, Serbia
anapantelic@imgge.bg.ac.rs

Proteins known as Late Embryogenesis Abundant Proteins (LEAPs) are intrinsically disordered and play a crucial role in desiccation tolerance. Resurrection plants can withstand prolonged desiccation and fully recover their metabolic function the next day after watering. We identified, structurally characterized *in silico,* and categorized LEAPs in hydrated and desiccated leaves of the ancient resurrection plant *Ramonda serbica* to get more insights into their physiological functions.
A representative LEA4 protein highly accumulated under water scarcity was recombinantly produced and purified. It was predicted to be a highly disordered and polar protein. Structural models of dimers obtained by AlfaFold2 served as input for molecular simulation dynamics (MDS). Further, coarse-grain MDS using the GROMACS simulation package showed a high propensity of LEA4 protein to form dimers mostly by hydrophobic interactions. After the production phase, the obtained trajectory was analyzed for chain-chain contact for both cases λ=1.06 and 1.10. The dissociation constants K_d were calculated and were $K_d \approx 0.000075$ M (75 μM) for the case λ=1.06 and $K_d \approx 0.0000758$ M (75.8 μM) for the case λ=1.10., for the protein concentration C_p=8.91*10-5 M.
The results obtained by size exclusion chromatography, dynamic light scattering, and atomic forced microscopy confirmed that the selected LEA4 protein is aggregation-prone. Our results are important for further elucidating protective LEAP's mechanism during desiccation. Structural flexibility and aggregation propensity of LEAP might be crucial in direct, or indirect (e.g. via LLPS) buffering free cellular water and maintaining the native conformations of biomolecules. We suggest that native or bioengineered LEAPs can be used to improve the drought resistance of crops.

Keywords: desiccation, intrinsically disordered proteins, late embryogenesis abundant proteins, molecular dynamic simulation.

Acknowledgement: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-66/2024-03/200042) and by the Science Fund of the Republic of Serbia-RS (PROMIS project LEAPSyn-SCI, # 6039663).

Pharmacogenomic landscape of Serbian population

Marina Jelovac*, Đorđe Pavlović, Biljana Stanković, Nikola Kotur, Vladimir Gašić, Bojan Ristivojević, Sonja Pavlović and Branka Zukić

Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
marina.jelovac@imgge.bg.ac.rs

Many drugs account for an extensive interpatient variability in dose requirement for achieving therapeutic benefit. Said variability is often largely influenced by common genetic variants in several pharmacogenes. Herein we present an overview of pharmacogenomic landscape of Serbian population on the example of 7 actionable pharmacogenes for which Clinical Pharmacogenetics Implementation Consortium issued guidelines for interpretation and use of haplotypes to estimate adequate dose in therapeutics for treatment of cancer, rheumatic and cardiovascular diseases: *TPMT, SLCO1B1, NUDT15, UGT1A1, VKORC1, CYP2C9* and *CYP4F2*.

A cohort of 113 patients from Serbia underwent whole exome sequencing. Generated FASTQ files were aligned to human reference genome build GRCh38 and variants were annotated using in-house pipeline. Calling of star alleles was performed using open-sourced bioinformatics tool Stargazer applied to BAM and VCF files. Subsequent analysis of frequency differences in Serbian versus European population was done using 1000 Genomes Project database and Chi square test in R program.

In 7 analyzed pharmacogenes 23 different haplotypes, denoted as star alleles, were found in total. Nine of these haplotypes had significantly different frequencies in Serbian comparing to European population – *CYP4F2**2, variant rs9923231 T in *VKORC1, TPMT**20 and *TPMT**43, *UGT1A1**36 and *UGT1A1**80, while in actionable pharmacogene *SLCO1B1* three haplotypes (*4, *20 and *37) with significantly higher frequencies in Serbian than in European population were detected.

Presented results show that frequencies of several important pharmacogenomics star alleles differ in Serbian comparing to European population. This could have serious clinical implications and potentially lead to more frequent development of adverse drug reactions when treating patients in need with standard drug doses. In Serbian population a potential for preemptive genotype testing exist when prescribing therapeutics for cancer (*TPMT, NUDT15, UGT1A1, SLCO1B1*) and rheumatic (*SLCO1B1, TPMT*) or cardiovascular diseases (*VKORC1, CYP2C9, CYP4F2*), and population-specific pharmacogenomics aspect should be taken into account for therapy optimization in clinical practice.

Keywords: population pharmacogenomics, bioinformatics, sequencing, biomedicine, precision medicine

Acknowledgement: This work has been funded by grant from the Ministry of Science, Technological Development and Innovation, Republic of Serbia (Grant No. 451-03-47/2023-01/ 200042) and EC project HORIZON-WIDERA-2021-ACCESS-02-01: PharmGenHUB (GA No. 101059870)

The morphological analysis of a Holter Electrocardiogram

M. Ćosić[1,*] and N. Miljković[2,3]

[1] Laboratory of Physics, Vinča Institute of Nuclear Sciences - National Institute of the Republic of Serbia, University of Belgrade, P.O. Box 522, Belgrade, Serbia
[2] University of Belgrade - School of Electrical Engineering, Belgrade, Serbia
[3] Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
mcosic@vinca.rs

We have developed a new method for analyzing the long scalar experimental data. By the time delay embedding, the oscillatory nature of the data was transformed into a sequence of loops of the system phase space trajectory. The proposed method reduces the classification of all possible loop types and rules of their succession.
The proposed method will be presented by analyzing the experimental data obtained from the Holter electrocardiogram of an arbitrarily chosen 79-year-old male patient and a 24-year-old healthy female individual. In the case of the patient, the system's trajectory consists of only 8 of the most simple loop types. Out of 64 possible loop transitions, only 49 play a significant role in the observed system dynamics. Further, we constructed a stochastic finite state machine capable of reproducing observed system trajectory statistically. We have observed many metamorphoses of the phase space distribution from clustered to branched.
The Holter electrocardiogram of a 24-year-old healthy individual showed no changes in the phase distribution shape, and the corresponding finite state machine has the same number of states but with radically different distributions of the transition probabilities.
Our method shows more sensitivity in detecting pathological states than standard heart rate variability assessments. Its topological nature makes it very robust to the disturbing effect of noise. Although, additional research on a larger experimental sample is required before the usefulness of the proposed morphological approach can be rightfully assessed. Results presented for just the two subjects clearly show the great potential of the dynamist approach aided with constructing the finite state machine.

Keywords: Poincaré plot, morphological method, heart rate variability, RR tachogram

Acknowledgement: The research was funded by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia (grant Nos. 451-03-47/2023-01/ 200017 and 451-03-66/2024-03/ 200017).

Identification of Synonymous Genetic Variants Associated with Idiopathic Thrombosis using whole exome sequencing

Martina Mia Mitić[1], Dušan Ušjak[1], Mirjana Kovač[2], Marija Cumbo[1], Sofija Dunjić Manevski[1], Branko Tomić[1] and Valentina Đorđević[1,*]

[1] Institute of Molecular Genetics and Genetic Engineering, Belgrade, Serbia
[2] Blood Transfusion Institute of Serbia, Belgrade Serbia
martinamiamitic@imgge.bg.ac.rs

The etiology of substantial number of thrombosis cases remains undetermined. Sequencing and comprehensive analysis of the entire exome in patients with idiopathic thrombosis enable the identification of novel gene variants potentially significant for disease onset and the discovery of previously unknown molecular mechanisms underlying this multifactorial condition.

This study aimed to investigate synonymous genetic variants potentially linked to thrombosis. While previous studies predominantly focused on missense and stop gain variants, the role of synonymous variants in thrombosis susceptibility remains relatively overlooked.

The study included 50 subjects: 17 patients with recurrent idiopathic thrombosis and 33 controls from the general population. Whole exome sequencing (WES) was performed according to protocol of the Beijing Genomics Institute and subsequent FASTQ files were processed to obtain Variant Call Format (VCF) files with annotations. While the entire exome was examined, a panel of 55 genes was selected for more focused analysis. The data were filtered to extract only synonymous variants within the genes in this panel.

Statistical analyses were performed by creating contingency table counts in R and by applying weights and utilizing chi-square or Fisher's exact tests in Python. A p value of $p < 0.05$ was defined as statistically significant.

A total of 15 synonymous variants exhibiting statistical significance in the idiopathic thrombosis group were identified. Variants were detected in PLG, PROC, ABO, KNG1, ADAMTS13, ACE, HIVEP1 and F2 genes.

In conclusion, although the study exhibits considerable limitations due to the small sample size, the identification of significant synonymous variants underscored a potential association between thrombosis and these genetic determinants. Further studies with larger cohorts are imperative to validate and expand upon these initial findings.

Keywords: Thrombosis, Synonymous variants, Whole exome sequencing, Idiopathic thrombosis

Machine learning approach for risk factors detection of pancreatic fistula and AI diagnostic systems development

Mikhail Potievskiy*, Sergei Ivanov, Andrei Kaprin, Ruslan Moshurov, Leonid Petrov, Peter Shegai, Pavel Sokolov and Vladimir Trifanov

FSBI "National Medical Research Radiological Centre" of the Ministry of Health of the Russian Federation, Kaluyhskaya region, Obninsk, Russia
potievskiymikhail@gmail.com

The aim of the study was to develop a predictive ML model for postoperative pancreatic fistula and to determine the main risk factors of the complication.
We performed a single-centre retrospective clinical study. 150 patients, who underwent pancreatoduodenal resection in FSBI NMRRC, were included. We developed ML models of biochemic leak and fistula B/C development. Logistic regression, Random forest and CatBoost algorithms were employed. The risk factors were evaluated basing on the most accurate model, roc auc, and Kendall correlation, $p<0.05$.
We detected a significant positive correlation between blood and drain amylase level increase in association with biochemical leak and fistula B/C. The CatBoost algorithm was the most accurate, roc auc 74%-86%. The main pre- and intraoperative prognostic factors of all the fistulas were tumor vascular invasion, age and BMI, roc auc 70%. Specific fistula B/C factors were the same. Basing on the 3-5 days data, biochemical leak and fistula B/C risk factors were blood and drain amylase levels, blood leukocytes, roc auc 86% and 75%.
We developed sufficient quality ML models of postoperative pancreatic fistulas. Blood and drain amylase level increase, tumor vascular invasion, age and BMI were the major risk factors of further fistula B/C development.

Keywords: machine learning, precision oncology, risk factor detection, pancreatoduodenal resection, pancreatic fistula

Viral presence in the 1000 Genomes Project data

Milana Djonovic[1,*] and Alexej Abyzov[2]

[1] Mayo Clinic College of Medicine and Science, Rochester, MN 55905, USA
[2] Department of Quantitative Health Sciences, Center for Individualized Medicine, Mayo Clinic, Rochester, MN 55905, USA
djonovic.milana@mayo.edu

This study focuses on detecting natural viruses in whole-genome sequencing (WGS) data obtained as part of the 1000 Genomes Project consortium. Through the use of alignment tools, filtering reads based on their quality and uniqueness of mapping to the viral reference, and assessing coverage across the viral genome, the study aimed to achieve reliable and sensitive virus detection. Analysis of the dataset revealed the presence of natural Epstein-Barr virus (EBV) in at least 12 samples out of 2504, which is distinct from its artificial counterpart commonly used for cell line transformation. These findings are consistent with the earlier results obtained on a smaller dataset of 750 samples within the same project. In addition to natural EBV, we identified human betaherpesvirus 6A, herpesvirus 6B, herpesvirus 7, herpesvirus 4, and T-lymphotropic virus 1 in 19 samples. This study demonstrated that viruses can be detected in WGS data, and that our methodology could be applied to healthy human tissues as well. By shedding light on the presence of viruses in healthy human tissues, this research could have important implications for personalized medicine and public health initiatives.

Keywords: bioinformatics, virus detection, 1kGP, WGS, EBV

Exploring biotechnological potential of LLDPE- and mixed plastics-degrading bacteria from contaminated soils

Milica Ciric[1,*], Clémence Budin[2], Tjalf de Boer[2], Brana Pantelic[1] and Jasmina Nikodinovic-Runic[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[2] Microlife Solutions, Amsterdam, The Netherlands
milica.ciric@imgge.bg.ac.rs

Global research efforts to develop biocatalytic processes based on microbial enzymes to alleviate plastic pollution are underway. Despite a proposed overlap between enzymatic capacity to degrade plastics and lignocellulosic biomass, bioprospecting efforts to identify microorganisms capable of degrading both types of substrates remain limited.
Plastic and lignocellulose degrading potential of 15 bacterial isolates from polluted Serbian soils, belonging to genera detected with abundance >1% in virgin plastics (LLDPE)- or post-consumer representative mixed plastics (MS)-enriched 16S metagenomes, was explored. Plastic polymers Impranil® DLN-SD (SD) and DL2077 (DL), bis(2-hydroxyethyl) terephthalate (BHET), polycaprolactone diol (PCL) and polylactic acid (PLA) and lignocellulosic polymers carboxymethyl cellulose (CMC), arabinoxylan (AXYL) and lignin (LIG) were provided as sole carbon source for the isolates.
16S rDNA dendrogram of 9 LLDPE-enriched isolates from genera Lysinibacillus, Rhodococcus, Hydrogenophaga, Pseudomonas, Nocardoides and Psychrobacillus and 6 MS-enriched isolates from genera Pseudomonas, Advenella and Paenibacillus showed no clear grouping, suggesting that relatively distinct isolates were selected for the screening. No zones of clearing, indicating complete substrate degradation is taking place, were observed. Single Advenella isolate demonstrated growth on all plastic substrates, while 11 isolates demonstrated growth on PCL, 10 on BHET, 6 on SD, 4 on DL and 4 on PLA. Fourteen isolates grew on all tested lignocellulosic substrates.
Genome mining of 3 sequenced isolates using PAZy database identified putative PU-, PET, PCL- and PLA-active enzymes in both Hydrogenophaga, growing on all plastic substrates except PLA and Pseudomonas, growing on BHET and PCL, while Lysinibacillus, with predicted PCL and PLA activities, demonstrated no growth on any of the tested plastic substrates. Lignocelullolytic enzymes (GHs, CEs, PLs and AAs) were also predicted in these isolates, demonstrating growth on all tested lignocellulosic substrates, using CAZy database.
Investigated isolates should be further explored using a wider range of plastic substrates and screening conditions.

Keywords: soil, 16S metagenomics, genome mining, PAZy, CAZy

Acknowledgement: This work was supported by the EU H2020 Research and Innovation Programe (grant agreement No. 870292, BioICEP) and by the Ministry of Science, Innovation and Technological Development of the Republic of Serbia (agreement No. 451-03-66/2024-03/200042). 16S rDNA sequences are deposited in the NCBI GeneBank database (accession numbers: PP594182-PP594194). 16S metagenomic and genome sequences are available upon request.

Improvement of PBMC Cell Types Classification in Healthy Samples

Minjie Lyu[1], Lin Xin[1], Lou T. Chitkushev[2], Guanglan Zhang[2],
Derin B. Keskin[3] and Vladimir Brusic[1,*]

[1] University of Nottingham, Ningbo, China
[2] Boston University, Boston, United States
[3] Dana-Farber Cancer Institute, Boston, United States
vladimir.brusic@nottingham.edu.cn

Peripheral blood mononuclear cells (PBMC) are used in the study of the immune system, infectious diseases, and vaccine development. PBMC are composed of six major cell types: B cells, T cells, Natural Killer cells, monocytes, classical dendritic cells, and plasmacytoid dendritic cells. Single cell transcriptomics (SCT) is an emerging technology that concurrently measures gene expression from tens or even hundreds of thousands of individual cells. It provides a higher resolution of gene expression measurement than traditional bulk-sequencing. 10x Genomics is a SCT technology that is able to capture more than 100,000 cells in a single study. Labelling cell types is the first and crucial step in most SCT studies. However, it is impractical to label more than ten thousand cells manually. Supervised machine learning methods such as artificial neural networks (ANN) are suitable for cell type classification. However, the proposed state-of-art accuracy of PBMC classification was less than 80%, which makes it undesirable to label PBMC from new datasets with the latest methods reaching the accuracy of 95% or more. Most classification errors are caused by using the datasets where cell types are incorrectly labelled.
We collected datasets from 10x genomics demonstration databases, including 28 PBMC datasets generated from healthy donors. Datasets were standardized into a common gene list containing 30,698 genes. Quality control (QC) was performed to eliminate dead or low-quality cells, and more than 250,000 cells passed our QC metrics. Four prediction methods (ANN classification, ANN super-class classification, profile-based prediction, and protein-marker-based prediction) were combined to label the cell types. To evaluate the results, we trained ANN models using data from one dataset and tested using the remaining 27 datasets. The average classification accuracy was 98.46%. Datasets with high-accuracy cell type labels can then be used for high-accuracy healthy PBMC cell type classification.

Keywords: cell type prediction, peripheral blood mononuclear cells, scRNA-seq, supervised machine learning, reference datasets

Integrating functional screening and bioinformatics for personalized medicine approaches in NSCLC

Miodrag Dragoj[1,*], Jelena Dinić[1], Sofija Jovanović Stojanov[1], Ana Stepanović[1], Ema Lupšić[1], Milica Pajović[1], Thomas Mohr[2], Sofija Glumac[3,4], Dragana Marić[4,5], Maja Ercegovac[4], Ana Podolski-Renić[1] and Milica Pešić[1]

[1] Institute for Biological Research "Siniša Stanković" - National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia
[2] Center for Cancer Research and Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria
[3] Institute of Pathology, University of Belgrade - School of Medicine, Belgrade, Serbia
[4] University of Belgrade - School of Medicine, Belgrade, Serbia
[5] Clinic for Pulmonology, University Clinical Center of Serbia, Belgrade, Serbia
miodrag.dragoj@ibiss.bg.ac.rs

Understanding the impact of tyrosine kinase inhibitors (TKIs) on multidrug resistance (MDR) in non-small cell lung carcinoma (NSCLC) is critical for effective cancer treatment. While TKIs target specific signalling pathways in cancer cells, they can also be substrates for ABC transporters and thus potentially induce MDR. This study aimed to assess the responses of 17 patient-derived NSCLC cultures to 10 widely used TKIs and to investigate the relationship between these responses and patient-specific mutational profiles using bioinformatics approaches.

An ex vivo immunofluorescence assay was used to determine the expression levels of the MDR markers ABCB1, ABCC1 and ABCG2. These expression data were then integrated with detailed genetic profiles obtained by next-generation sequencing (NGS) and analysed using bioinformatics tools to identify correlations and functional significance. The results showed a differential response of NSCLC cultures to the TKIs, with erlotinib showing significant efficacy regardless of mutational burden or EGFR status. However, erlotinib also led to an increase in ABCG2 expression, highlighting the complex challenge of MDR in treatment.

Gene set enrichment analysis showed that genetic alterations in key signalling pathways, including ErbB, Ras and PI3K-Akt, as well as mechanisms associated with EGFR-TKI resistance, likely influenced the differential responses to TKIs. Although no consistent patterns were observed between ABC transporter expression and genetic alterations, our integrative bioinformatics approach provided valuable insights into the mechanisms underlying TKI resistance and sensitivity.

These results emphasise the critical role of bioinformatics and functional sensitivity screening in addition to mutational profiling for the selection of appropriate TKI treatments. Our study highlights the potential of personalised medicine in NSCLC therapy by taking into account drug sensitivity, off-target effects, MDR risks and patient-specific genetic landscapes.

Keywords: lung cancer, genomics, multidrug resistance, tyrosine kinase inhibitors, targeted therapy

Acknowledgement: This research was supported by the Science Fund of the Republic of Serbia, #7739737, Functional diagnostics in non-small cell lung carcinoma - a new concept for the improvement of personalized therapy in Serbian patients - TargetedResponse.

Two different methods to assess PSA-test results in patients with prostate cancer

Nenad Vesić* and Andjelka Hedrih

Mathematical Institute of Serbian Academy of Sciences and Arts, Belgrade, Serbia
n.o.vesic@outlook.com

Prostate cancer is the third most common tumor in men, with a 5-year relative survival rate of 34% for distant SEER stage cases. This makes prostate cancer particularly suitable for estimating the optimal set of clinical parameters essential for monitoring tumor progression or regression post-treatment.
In this study, PSA test results from patients of different ages and races with the same type of prostate cancer were analyzed. Two methods were applied: the method of evaluating established results using exact measures (as shown in Vesić, Mačukanović-Golubović, Ilić, 2017), and a statistical method utilizing mean values. These methods were applied to data from twenty prostate cancer patients taken from the End Results (SEER) Program, as well as to different sub-populations categorized by age and race. The results of these methods were compared, highlighting the differences between statistical analysis outcomes and those obtained through exact measures evaluation.
Additionally, descriptive statistics were performed on a population of 1,187,083 prostate cancer patients from the SEER Program. This included frequency distributions for patient age, race, histological tumor type, chemotherapy status, time from diagnosis to therapy initiation, and correlations between race and tumor histology, as well as tumor histology and living area. A Cox regression model was applied to the entire examined population to identify variables related to survival rates. The initial variables included in the model were: Gleason score, PSA value, median household income adjusted for 2022 inflation, rural or urban residence, age, and race.

Keywords: big data analysis, prostate cancer, PSA test, Cox regression, exact measures method.

Acknowledgement: We are grateful to the financial support from Ministry of Science, Technological Development and Innovation of Republic of Serbia trough Mathematical Institute of Serbian Academy of Sciences and Arts.

Sequence-based Hierarchical Classification of Tandem Repeats using Neural Network Models

Nevena Ćirić* and Jovana Kovačević

Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
nevena.ciric@matf.bg.ac.rs

Repeat proteins are a widespread class of mostly non-globular proteins containing repetitive subsequences, a so-called repeat units that often occur in tandem arrangements when observed in 3D structure of the protein. These tandem repeats are considerably diverse, ranging from the repetition of a single amino acid to domains of 100 or more residues.
Improvement of the methods for identification of protein tandem repeats and subsequently the increasing number of the known proteins containing repetitive elements necessitates their classification to facilitate further understanding of their sequence-structure-function relationships. According to Kajava's classification scheme based on the repeat unit's length, general structural arrangement and mode of interaction between the repeat units [1], tandem repeats are classified into five main classes and further divided into subclasses that reflect repeat unit topology, differing in secondary structure arrangement and/or overall structure within the repeat.
The classical approach to obtain the (sub)class assignment for a newly identified tandem repeat is by simply transferring this information from the "master" repeat unit, that is the repeat unit from database of predetermined tandem repeats with associated (sub) classes found to be most similar to the newly identified tandem repeat. This procedure usually implies some kind of structural search algorithm in order to assign master repeat unit, which further implies known tertiary structure of the protein with newly discovered tandem repeat. With intention to tackle the problem of analyzing proteins with unknown 3D structure and facilitate classification of tandem repeats by using only the sequence information, as well as to explore sequence-structure relationship between repeat units sequence and structural characteristics of their corresponding (sub)class, here we propose neural network based model for classification of tandem repeats based on the multiple sequence alignment of its units sequences. Additionally, this model can be further utilized to create an end-to-end pipeline for identification and classification of tandem repeats only from sequence information.

Keywords: tandem repeats, classification, neural network, sequence-based

References

1. 10.1016/j.jsb.2011.08.009

Analysis of SARS-CoV-2 Variant Sequences for Identification of Unique Insertions and Deletions for qPCR Detection of Emerging Variants

Yolshin Nikita*, Varchenko Kirill, Komissarov Andrey and Lioznov Dmitry

Smorodintsev Research Institute of Influenza, Saint-Petersburg, Russia
Nikita.yolshin@gmail.com

The global COVID-19 pandemic, which began in 2020, has claimed the lives of over 7 million people by conservative estimates. It is now evident that the novel coronavirus will become a seasonal virus, yet it continues to evolve, with antigenically novel and more contagious variants triggering new waves of the epidemic. Monitoring the emergence and spread of new variants is crucial, as the findings of such research impact decision-making, vaccine development, therapeutic strategies, and their implementation. Polymerase Chain Reaction (PCR) offers a rapid and cost-effective solution to this challenge.

Many significant variants' genomes carry unique insertions and deletions, which have proven to be convenient targets for developing RT-PCR assays. Our in silico work began in the early days of concerning variant emergence, with detection assays ready by the time new SARS-CoV-2 variants were imported into the country. To select optimal targets—unique insertions and deletions for variant-specific assays—all relevant variant sequences from the GISAID database were aligned using the MAFFT program. Relative substitution frequencies for each lineage were calculated based on amino acid substitution and insertion/deletion annotations using Python language and Pandas and Numpy libraries. Subsequently, multiple oligonucleotide sets were designed, and the most specific, sensitive, and effective PCR assay was chosen for implementation.

Unique insertions and deletions were proposed for the detection of SARS-CoV-2 lineages and variants: a deletion in the ORF1ab gene (SGF3675del, 11288–11296) for Alpha lineage detection, an insertion in the S gene (insN679KGIAL, 23573-23584) for AT.1 lineage, a deletion in the S gene (156-157EFdel, 22029-22034) for Delta lineage, an insertion in the S gene (Ins214EPE, 22205-22213) for the Omicron BA.1 subvariant, and a deletion in the N gene (ERS31del, 28338-28346) for Omicron lineage (all subvariants).

The developed assays may prove useful in the future for detecting yet-to-exist variants or determining the future seasonal SARS-CoV-2 or its variants.

Keywords: SARS-CoV-2, qPCR, mutations, insertions, deletions

Analysis of Influenza Virus Sequences in Russia for the Current Epidemic Season 2023-2024

Yolshin Nikita*, Komissarov Andrey and Lioznov Dmitry

Smorodintsev Research Institute of Influenza, Saint-Petersburg, Russia
Nikita.yolshin@gmail.com

During the current epidemic season 2023-2024, approximately 1700 influenza viruses were sequenced at the Smorodintsev Influenza Research Institute from patients with acute respiratory viral infections from 20 regions of the Russian Federation. Next-generation sequencing technologies (Illumina NextSeq 2000, MGI DNBSEQ-G400) were utilized for sequencing. Genome assembly onto the reference was performed using bwa-mem2 2.2.1 and samtools 1.6 programs, with consensus generation by the ivar 1.3.2 tool. The tool fastqc 0.12.1 was used to assess the quality of data and trimmomatic 0.39 to trim the reads.

Genome alignment of the current season's viruses with earlier variants revealed notable characteristics. The vast majority (99%) of influenza viruses sequenced in Russia belonged to type H3N2, clade 3c.2a1b.2a.2a.3a.1 (vaccine virus A/Thailand/8/2022-like), sharing substitutions E50K, I140K (antigenic site A), and I223V (antigenic site D). Within clade 2a.3a.1, a subgroup with N122D substitution (loss of a potential glycosylation site) was predominant, with 85% representation among Russian viruses. Notably, viruses within this subgroup commonly featured K276E substitution at HA antigenic site C.

Additionally, several influenza A H1N1 and influenza B samples were successfully sequenced during the season. All H1N1 samples fell within clade 6B.1A.5a.2a (A/Victoria/2570/2019 vaccine virus clade), with characteristic amino acid substitutions in HA, including K130N, N156K, L161I, V250A, and E506D. These viruses also encoded amino acid substitutions K54Q, A186T, E224A, R259K, and K308R in HA1, characteristic of 5a.2a viruses. Notably, 5a.2a.1 clade viruses and A/Sydney/5/2021-like viruses were not detected in Russia during this season. Several distinct groups of Russian isolates were identified within the 5a.2a clade, exhibiting specific amino acid substitutions such as T120A and K169Q (antigenic site Ca1), V321I, and P137S (A/Netherlands/10468/2023-like).

Influenza surveillance plays a critical role in monitoring and understanding the dynamic evolution of influenza viruses, aiding in the timely detection of emerging strains and informing public health measures to mitigate the impact of seasonal epidemics and potential pandemics.

Keywords: Influenza, H3N2, H1N1, surveillance, NGS

A Graphical User Interface for Automated BLAST Analysis and Phylogenetic Tree Construction

Nikola Đorđević[1,2], Ivan Skadrić[3], Slaviša Stanković[4], Zorica Knežević-Jugović[2] and Snežana Đorđević[1]

[1] Agrounik ltd, Belgrade, Serbia
[2] University of Belgrade - Faculty of Technology and Metallurgy
[3] Klaren ltd, Belgrade, Serbia
[4] University of Belgrade - Faculty of Biology
nikola.djordjevic@agrounik.rs

To enhance modern laboratory workflow, we developed software with a graphical user interface (GUI) that automates BLAST analysis, creates, and displays phylogenetic trees for multiple sequences in FASTA format. This software features time-saving functionality that increases productivity in the identification of microorganisms based on both 16S and ITS analysis and matches them in reference databases on a local server.

Both the GUI and the software backend are written in Python and incorporate two key functions, perform_blast_analysis and create_phylogenetic_tree, which utilize methods from the BioPython library. The perform_blast_analysis function automates and optimizes BLAST analysis by utilizing methods from the Bio Blast subpackage to execute queries for multiple sequences and generate results in XML output format. Once generated, the results can be extracted from XML and saved as a CSV file, providing a detailed report containing information on the five best matching sequences, including query ID, sequence name, topic ID, bacterial type, identification percentage, alignment length, E-value, and score.

The create_phylogenetic_tree function automates phylogenetic analysis and visualization, starting with sequence alignment by the Muscle tool from the Bio Align subpackage. Genetic distance is estimated using methods from the Bio Phylo package with an 'identity' model, and the phylogenetic tree is constructed by the UPGMA algorithm. Visualization of the phylogenetic tree is performed using the matplotlib package, providing the user with a clear view of the genetic lineages. By integrating these steps, the application significantly speeds up screening analysis and automates the complex processes of microbial taxonomy determination and examination of evolutionary relationships based on DNA sequencing.

This application highlights the importance of intuitively developing a GUI with functions tailored for specific laboratory tasks, based on the integration of laboratory, bioinformatics, and software development experience. By implementing this software into laboratory workflows, a significant reduction in the time required to process FASTA files with multiple entries can be achieved.

Keywords: BLAST, phylogenetic tree construction, microbial identification

DNA Mechanics peculiarities: Model for Twist and Stretch Coupling

P. P. Kanevska and S. N. Volkov

Bogolyubov Institute for Theoretical Physics of the
National Academy of Sciences of Ukraine, Kyiv 03143, Ukraine
snvolkov@bitp.kiev.ua

Our study introduces a model of DNA macromolecule deformation, which integrates both external and internal deformation components with their interrelations. The external deformation includes the twist and stretch of the DNA double helix, while the internal deformation involves the relative shift of base pairs within the double helix. The model incorporates the coupling between the twist and stretch of the double helix and with its internal restructuring. The model reveals counterintuitive dynamics: under stretching forces, the DNA double helix can exhibit increasing in twist up to a critical force. Over this force, the double helix behaves more like a conventional filament, showing untwisting in response to stretching. These findings highlight the significant impact of coupling between internal and external components on the mechanical behaviour of the double helix when subjected to forces lower of the critical point. The agreement of our model results with experimental data also confirms its accuracy and applicability.

Keywords: DNA mechanics, twist-stretch coupling, molecular modelling, structural dynamics, biophysics

Acknowledgement: I would like to express my gratitude to the University of Exeter for the opportunity to conduct research as an Honorary member. Additionally, I would like to acknowledge the support from the British Academy and Cara (the Council for At-Risk Academics) under the 'Researchers at Risk' programme, which is funded primarily by the UK Government.

Enhancing Immunogenicity Assessment of C57BL/6 T-cell Epitopes

Zitian Zhen[1], Alexis A. Howard[2], Derin B. Keskin[1,2], Vladimir Brusic[1,3], Lou Chitkushev[1,*] and Guang Lan Zhang[1]

[1] Metropolitan College, Boston University, Boston MA, USA
[2] Dana-Farber Cancer Institute, Boston MA, USA
[3] University of Nottingham Ningbo, Ningbo, China
ltc@bu.edu

The C57 Black 6 (C57BL/6) mice, recognized for their genetic uniformity, are among the most widely utilized inbred laboratory animals in immunology research and vaccine development. We propose developing a bioinformatics system for the *in silico* prediction of MHC class I restricted T-cell epitopes in C57BL/6 mice. Among the multiple steps of the MHC class I antigen processing and recognition pathway, MHC binding is considered the most selective step in T cell recognition. However, many bioinformatics systems focus solely on modeling MHC binding to predict binders, which leads to higher rates of false positives. Recent technological advancements in mass spectrometry (MS) have provided abundant MHC class I ligand data, allowing the incorporation of antigen processing steps. We collected >5,000 H2-D^b and >5,000 H2-K^b binding peptides, along with >4,000 and >5,000 eluted ligands from public databases, respectively. Additionally, thermostability assessment of peptide-MHC binding is crucial for accurate immunogenicity predictions, as stability affects antigen presentation efficiency and T cell activation. We utilized data generated from Dana-Farber Cancer Institute's temperature gradient experiments, which yielded >3,000 H2-D^b and >5,000 H2-K^b binding peptides. Our work integrates these factors into a computational system for epitope identification in C57BL mice. Utilizing deep learning methods, we trained and validated epitope prediction models using natural ligands and thermostability models using the proprietary data generated by our collaborators. We compared the performance of our models with existing prediction tools validated by many benchmark studies. Our integrated model exhibited superior overall predictive capabilities.

Keywords: bioinformatics system, deep learning, T-cell epitope, MHC binding, C57BL/6

Acknowledgement: DBK would like to acknowledge support from R01-HL157174 and NIH/NCI 3UG1CA189955-09S1

Trends for Artificial Intelligence, Machine Learning, and Deep Learning applications in plant breeding

Eftekhari M.[1], Ma C.[2] and Yuriy L. Orlov[3,4,5,*]

[1] Department of Horticultural Sciences, Faculty of Agriculture, Tarbiat Modares University, Tehran, Iran
[2] Center of Bioinformatics, College of Life Sciences, Northwest A&F University, Shaanxi, China
[3] Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia
[4] Agrarian and Technological Institute, Peoples' Friendship University of Russia, Moscow, Russia
[5] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
m.eftekhari@modares.ac.ir
chuangma2006@gmail.com
orlov@d-health.institute

The field of plant breeding has witnessed a paradigm shift driven by advancements in artificial intelligence (AI) technologies, including machine learning (ML) and deep learning (DL) technologies. These cutting-edge techniques have transformed our understanding of plant biology, reshaping the landscape of plant breeding. AI-assisted omics techniques offer insights into plant-pathogen interactions and facilitate the identification of stress-responsive genes.
We have organized thematic journal issue at Frontiers in Plant Science – Research Topic "Applications of artificial intelligence, machine learning, and deep learning in plant breeding". Here we present the trends of AI applications and the prediction problems challenged in plant science. We collected research papers on the topic of AI applications in plant biology in areas of sequencing data analysis, image recognition, technology process optimization. It includes description of the potential of AI algorithms, particularly ML and DL, in decoding complex omics data to elucidate the molecular foundations of plant defense. Climate change poses significant threats to agricultural systems, emphasizing the importance of elucidating cold defense mechanisms in crops. From the methodical point of view, the Self Organizing Maps (SOM)-based ML methods can decipher gene expression patterns in response to different temperature regimes. The YOLO (You Only Look Once) architecture, known for its real-time object detection capabilities, is employed for object detection in plant image analysis. The efficacy of the CFNet-VoVGCSP-LSKNet-YOLOv8s model in accurately identifying cotton pests and diseases amidst challenging environmental conditions is shown. Another key phenotypic trait in plants - pubescence, correlates with stress resistance, particularly in wheat. Overall, AI, ML, and DL techniques offer unique opportunities from deciphering complex omics data to automating phenotypic trait analysis and disease detection to revolutionize breeding practices, develop stress-tolerant and high-yielding crop varieties, and contribute to global food security in the face of escalating environmental challenges.

Keywords: agrobiology, AI, Machine Learning, omics data

Acknowledgements: The study is supported by Russian Science Foundation (grant project 23-44-00030).

GWAS study for severe COVID-19 linked with thromboinflammation syndrome

Olga Y. Bushueva[1,*], Alexey V. Loktionov[1] and Yuriy L. Orlov[2,*]

[1] Research Institute for Genetic and Molecular Epidemiology, Kursk State Medical University, Kursk, Russia
[2] Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow 19991, Russia
olga.bushueva@inbox.ru
orlov@d-health.institute

COVID-19 pandemic raised challenges for medical statistics studying genome associations and nucleotide polymorphism as predispostion to the disease. We studied association of COVID-19 with thromboinflammation. The aim of the study was to replicate associations of GWAS-significant loci with severe COVID-19 in the population of Central Russia, to investigate associations of the SNPs with thromboinflammation parameters, to analyze gene-gene and gene-environmental interactions.
DNA samples from 798 unrelated Russian subjects (199 hospitalized COVID-19 patients and 599 controls with a mild or asymptomatic course of COVID-19) were genotyped using probe-based PCR for 10 GWAS-significant SNPs: rs143334143 CCHCR1, rs111837807 CCHCR1, rs17078346 SLC6A20-LLZTFL1, rs17713054 SLC6A20-LLZTFL1, rs7949972 ELF5, rs61882275 ELF5, rs12585036 ATP11A, rs67579710 THBS3, THBS3-AS1, rs12610495 DPP9, rs9636867 IFNAR2. SNP rs17713054 SLC6A20-LZTFL1 was associated with increased risk of severe COVID-19 in the entire group (risk allele A, OR = 1.78, 95% CI = 1.22–2.6, p = 0.003), obese individuals (OR = 2.31), patients with low fruit and vegetable intake (OR = 1.72), low physical activity (OR = 1.93), and nonsmokers (OR = 1.65). This SNP correlated with increased BMI (body mass index) (p = 0.006) and worsened thrombodynamic parameters, delayed appearance of spontaneous clots. SNP rs17078346 SLC6A20-LZTFL1 was linked with increased BMI (p = 0.01) and severe COVID-19 in obese individuals. The pairwise combination rs7949972 ELF5-rs61882275 ELF5 was a priority in determining susceptibility to severe COVID-19 (it was included in all the most significant SNP-SNP interaction models and described 4.04% of the entropy of severe COVID-19). Overall, this study represents a comprehensive molecular-genetic and bioinformatics analysis of the involvement of GWAS-significant loci in the molecular mechanisms of severe COVID-19, gene-gene and gene-environmental interactions, and provides evidence of their relationship with thromboinflammation parameters in patients hospitalized in intensive care units.

Keywords: chronic diseases, COVID-19, genotyping, SNP, GWAS, thromboinflammation syndrome

Integration of bioinformatics data for crop plant breeding

Yuriy L. Orlov[1,2,*], Haoyu Chao[3], Shilong Zhang[3],
Vladimir A. Ivanisenko[2] and Ming Chen[3]

[1] Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow 19991, Russia
[2] Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, Novosibirsk 630090, Russia
[3] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China
orlov@d-health.institute

In agrobiology the crop plant breeding involves selecting new plant varieties with desirable traits such as increased yield, improved disease resistance, and enhanced nutritional value. In recent years, the emergence of high-throughput omics technologies has revolutionized crop plant breeding by providing vast amounts of data on the molecular mechanisms underlying plant development, and responses to environmental stresses. However, to effectively use these technologies, integration of multi-omics data from different databases is required. Another important aspect of biotechnology is the growth of AI and Machine Learning applications. Integration of omics data provides a comprehensive understanding of the biological processes underlying plant traits and their interactions. The number of sequenced crop genomes has continued to rapidly grow in recent years, providing valuable resources for agricultural research. Additionally, epigenomics and transcriptomics have become increasingly important in crop breeding, providing insights into gene regulation and aiding in the identification of desirable traits. The SRA database has seen a continuous increase in epigenomic and transcriptomic data, further emphasizing the significance of these fields for crop breeding. Proteomics and metabolomics have continued to develop in crop breeding, allowing for a deeper understanding of plant molecular mechanisms. Several omics databases have been developed to store and analyze large-scale omics data for different crop species, including rice, maize, wheat, and soybean. These databases provide a wealth of information on the genetic makeup, epigenome regulation, gene expression profiles, protein functions, and metabolic pathways of crops, which can be used to improve breeding programs. The use of genomic databases such as NCBI Assembly, Genome Warehouse, EnsemblPlants, Phytozome, and PlantGDB provides access to genome sequences, gene annotations, and functional annotations for many crop species, including rice, maize, soybean, wheat. We highlights the importance of integrating omics databases in crop plant breeding, discusses available omics data and databases, describes integration challenges, and highlights recent developments and potential benefits. Taken together, the integration of omics databases is a critical step towards enhancing crop plant breeding.

Keywords: Omics, Data Integration, Databases, Plant biology, Crop plant breeding

The sequence complexity estimates: algorithms and applications

Yuriy L. Orlov[1,*] and Nina G. Orlova[2]

[1] Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia
[2] Financial University under the Government of Russian Federation, Moscow, Russia
orlov@d-health.institute

We discuss current methods and tools for algorithmic estimates of genetic texts (information and entropy measures). The search DNA regions with the extreme statistical characteristics is important for biophysical models of chromosome function and gene transcription regulation in genome scale. The complexity profiling has been applied to segmentation and delineation of genome sequences, search for genome repeats and transposable elements, applications to next-generation sequencing reads. We review the complexity methods and new applications fields: analysis of mutation hotspots loci, analysis of short sequencing reads with quality control, and alignment-free genome comparisons. The algorithms implementing various numerical measures of text complexity estimates including combinatorial and linguistic measures have been developed before genome sequencing era. The series of tools to estimate sequence complexity use compression approaches, mainly by modification of Lempel-Ziv compression. Most of the tools available online provide service for whole genome analysis. Novel machine learning applications for classification of complete genome sequences also include sequence compression and complexity algorithms. We present comparison of the complexity methods on the different sequence sets. Further, we discuss approaches and application of sequence complexity for proteins. The complexity measures for amino acid sequences could be calculated by the same entropy and compression-based algorithms. But the functional and evolutionary roles of low complexity regions in protein have specific features differing from DNA. The tools for protein sequence complexity aimed for protein structural constraints. Low complexity regions in protein sequences are conservative in evolution and have important biological and structural functions. Finally, we summarize recent findings in large scale genome complexity comparison and applications for coronavirus genome analysis.

Keywords: algorithms, text complexity, entropy, Lempel-Ziv compression, genetic code, low complexity regions, sequencing artefacts, genomic rearrangement, alignment-free, genome comparison, online tools

Analysis of structural features of DNA in tRNA genes

Ekaterina A. Savina[1], Anastasia A. Anashkina[1], Irina A. Il'icheva[1] and Yuriy L. Orlov[2,*]

[1] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow, Russia
[2] Sechenov First Moscow State Medical University of the Russian Ministry of Health (Sechenov University), Moscow, Russia
orlov@d-health.institute

RNA polymerase III (Pol III) transcribes tRNA genes using type 2 promoters, namely intragenic boxes A and B. In addition, 5'-flanking regions of tRNA genes of plants and S. pombe contain octanucleotides similar to the well-known TATA-box in the promoters of mRNA genes. The TATA box has not been found in other eukaryotes, although TBP is the component of the transcription factor TFIIIB. Archaea use the orthologue. The goal of this work was to determine the position of possible TBP binding while tRNA genes transcription in various eukaryotes as well as the comparison of structural properties of nucleotide sequences of tRNA genes and their up-stream regions in eukaryotes, archaea and bacteria. We have analyzed representative sets of tRNA gene sequences from 11 organisms from the Genomic tRNA Database (GtRNAdb) (http://gtrnadb.ucsc.edu). The nature of the upstream DNA sequences has been discussed earlier. It was proposed to consider these areas as a fine-tuning control of the transcriptional activity and starting specificity of a given tRNA gene. The results of experimental work already existing by that time led to the conclusion that transcription initiation of eucaryotic tRNA genes thus appears to depend critically on DNA conformation, both within and around the genes. The characteristics of textual and spatial structure in the vicinity of the start of tRNA genes were detected. Textual characteristics of 60 bp sequences aligned to the ends of the genes were also analyzed and the position of B-box was defined. It was found that belonging to different domains has little effect on the consensuses of A and B boxes. They differ only at the degree of conservatism of some positions. The mechanical properties of the upstream regions of all eukaryotes allow TBP to bind not in one, but in several positions, but with different affinities. These properties are less pronounced in archaea, and absent in bacteria.

Keywords: tRNA transcription, TBP (TATA-binding protein), text complexity, archaea, bacteria, eukaryotes, DNA spatial structure

Versatile Multi-Sample Single Cell RNA-Seq Pipeline with Extensive Customization Options

Aleksandar Danicic*, Nevena Vukojicic,
Aleksandar Baburski and Ana Mijalkovic Lazic

Velsera, Belgrade, Serbia
aleksandar.danicic@velsera.com

Single-cell RNA sequencing (scRNA-seq) technology has become the state-of-the-art approach for describing cell subpopulation classification and cell heterogeneity. It allows addressing medical questions such as the role of rare cell populations contributing to disease progression and therapeutic resistance.
Presented here is the "Multi-Sample Clustering and Gene Marker Identification with Seurat 4.1.0", a highly customizable workflow for single-cell data analysis, implemented in the Common Workflow Language (CWL). The workflow consists of three steps: 1. Loading scRNA-seq Expression Datasets, 2. Quality Control and Preprocessing, and 3. Clustering and Identification of Gene Markers. It supports gene-cell count matrices generated by several commonly used quantifiers (Cell Ranger counts, STAR solo, Salmon Alevin, Kallisto BUStools) coming from single or multiple single-cell datasets, different batches, as well as single or multiple samples combined in a single SingleCellExperiment object.
Each workflow step contains several implemented options, allowing a high level of customization. The quality control can be performed manually or automatically using several options for normalization (LogNormalize, Deconvolution, SCnorm and Linnorm) and batch effect correction (Seurat and Harmony). The workflow utilizes Seurat's graph-based approach for clustering, enabling the selection of multiple clustering resolutions. The identification of gene markers on a cluster level is performed by differential expression analysis step using various tests (wilcox, bimod, roc, and DESeq2).
To illustrate the utilization of this workflow in a standard single-cell analysis, two open access datasets containing cells isolated from human pancreatic cells were processed. Four clustering resolutions were employed to achieve different degrees of granularity, after which cluster-specific marker genes were identified.
The workflow is available on the Cancer Genomics Cloud (CGC), powered by Seven Bridges and funded by the NCI. CGC is a flexible cloud platform that ensures fast execution, scalability and reproducibility of the results, offering over 1000 bioinformatics workflows. To enable researchers to use this analysis as a guideline, this analysis was made as a public project.

Keywords: single-cell transcriptomics, cloud computing

Using Singular Value Decomposition for Extracting Underlying Gene Expression Patterns in Transcriptomic Analysis

Biljana Stankovic*, Mirjana Novkovic and Nikola Kotur

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Belgrade, Serbia
biljana.stankovic@imgge.bg.ac.rs

Singular Value Decomposition (SVD) is a mathematical approach that can be useful in analysis of transcriptome data. SVD is based on transforming gene expression data from genes × arrays to reduced eigengenes/eigenarrays vector space. In high-throughput analysis, the goal is often to reduce the dimensionality of data, excluding non-informative noise, and to extract patterns reflecting biological processes. Arranging the data based on eigenvectors provides a comprehensive overview of gene expression dynamics, where individual genes are classified into groups of similar regulation and function, or similar cellular state and biological phenotype.

Here we applied SVD to microarray gene expression data involving 21,176 genes from ulcerative colitis patients both glucocorticoid-sensitive (n=20) and glucocorticoid-resistant (n=20). Our aim was to validate SVD based methodology as an alternative to classical differential gene expression analysis (PMID: 20941359).

The SVD process involves decomposing a matrix $Am \times n$ (m=number of genes, n=number of patients) into three matrices $Um \times m$, $Dm \times n$, and $V^T n \times n$. The V^T matrix can be utilized to identify the eigengenes that differ the most between the sensitive and resistant patient groups. The greatest differences were found for eigengenes 7, 4, and 5 (p=0.007, p=0.078, and p=0.090, respectively (t-test)). For each eigengene of interest, lists of genes with the highest absolute values of projection and correlation were identified and used for gene and disease ontology analysis. Our results showed that the top 40 genes with the highest projection on selected eigengenes participate in the same five most important biological processes as the genes obtained from the differential gene expression analysis (PMID: 20941359).

In summary, SVD is a powerful tool for gene expression analysis, capable of isolating significant biological patterns. Further validation on additional datasets is necessary to confirm the robustness of SVD compared to more commonly used methods for differential expression analysis.

Keywords: differential gene expression, dimensionality reduction, eigengenes

Genome-wide association study identified genetic signal in cystatin genes associated with Long COVID-19

Marija Laban-Lazovic[1], Marko Zecevic[2,3], Nikola Kotur[2], Vladimir Gasic[2], Bojan Ristivojevic[2], Vesna Skodric-Trifunovic[1,4], Tatjana Adzic-Vukicevic[1,4], Branka Zukic[2], Sonja Pavlovic[2] and Biljana Stankovic[2,*]

[1] Clinic of Pulmonology, University Clinical Centre of Serbia, Belgrade, Serbia
[2] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
[3] Velsera, Belgrade, Serbia
[4] Faculty of Medicine, University of Belgrade, Serbia
biljana.stankovic@imgge.bg.ac.rs

Long COVID-19 is characterized by symptoms that persist for more than three months after the acute phase of a SARS-CoV-2 infection. It can affect individuals who had both severe and mild cases of COVID-19. The genetic predisposition to this condition remains largely unexplored.
This study aimed to identify genomic loci associated with Long COVID-19. The study included 92 patients with confirmed infection with SARS-CoV-2 during the delta variant wave and treated in hospital settings. These patients were monitored for up to six months after the acute infection. All patients were genotyped using the Illumina Infinium global screening array, which covers over 700,000 genomic variants. Imputation was employed to expand the number of variants to 12,001,939 using the 1kGP Phase 3 human reference panels. We conducted a genome-wide association analysis using an in-house built pipeline on the Cancer Genomic Cloud platform.
Our results identified association signals in the 20p11.21 genomic locus, with the lead variant being rs1275745396 ($p = 6.181x10-8$), located within cystatin genes (*CST3*, *CST4*, *CST1*). The cystatin family consists of cysteine protease inhibitors, involved in various physiological processes. Previous literature has linked serum and saliva cystatin levels to COVID-19 severity and outcomes. Moreover, cystatin's role in taste perception has been recognized previously, which is in line with the well-known perturbations in sensory perception caused by COVID-19. This study is the first to find a genetic association between cystatin genes and Long COVID-19. Further research is necessary to elucidate the role of the cystatin protein family and the biological processes underlying Long COVID-19. A deeper understanding of Long COVID-19 can inform the development of preventive and therapeutic strategies.

Keywords: GWAS, Long COVID-19, cystatins

Acknowledgment: This work was supported by the Ministry of Education, Science and Technological Development, Republic of Serbia (Grant No 451-03-66/2024-03/200042). 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 and 75N91019D00024.

Integration of Whole Exome and Single-Cell Transcriptomic Data Analysis to Identify Potentially Pathogenic Variants in Unicuspid Aortic Valve Disease

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

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[2] Institute for Cardiovascular Diseases Dedinje, Belgrade, Serbia
valentina@imgge.bg.ac.rs

Unicuspid aortic valve (UAV) disease is a congenital heart defect that can lead to severe cardiovascular complications. This study aimed to identify genetic variants contributing to UAV disease by integrating whole exome sequencing (WES) data with single-cell transcriptomics of the developing human heart.

WES was conducted on 28 subjects, including 9 with UAV and 19 non-affected family members by using protocol of the Beijing Genomics Institute (BGI). To refine the candidate gene set, previously published single-cell RNA sequencing data from 6.5-7 weeks post-conception (PCW) embryonic hearts were utilized [1]. The cell type and differential gene expression analyses were performed using Python, employing libraries such as Scanpy and scVI. WES data were filtered for high-impact and damaging missense variants, as predicted by three independent *in silico* tools, with an allele frequency of less than 10% in the GnomAD database, in genes that were notably expressed in cell types involved in aortic valve development. Subsequently, g:Profiler was utilized to perform functional profiling of the candidate genes and principal component analysis (PCA) was conducted to identify clustering patterns among the UAV patients.

The analysis identified 308 candidate variants in 283 genes, the majority of which are crucial in maintaining and organizing the extracellular matrix, supporting cellular adhesion and signaling, and contributing to the development of anatomical structures. Among these, 62 genes had damaging variants present in at least two UAV patients. Additionally, 15 novel variants were identified, not previously reported in the GnomAD database. Eleven variants were classified as pathogenic or likely pathogenic according to ClinVar or ACMG (American College of Medical Genetics and Genomics) criteria. Further, the PCA results revealed significant genetic variation across the UAV patients, with some patients showing closer proximity in affected gene profiles, suggesting potential clustering.

In conclusion, the approach of integrating WES data with existing single-cell transcriptomics data provided valuable insights into the genetic underpinnings of UAV disease. The study identified several novel candidate genes and variants, enhancing our understanding of the genetic basis of this congenital heart defect and potentially guiding future research, and diagnostic or therapeutic strategies.

Keywords: congenital heart disease, unicuspid aortic valves, genetic variants, whole exome sequencing, single-cell transcriptomics

1. 10.1016/j.cell.2019.11.025

Identifying the cluster of differentiation markers deregulated in colon cancer through analysis of Gene Expression Omnibus database

Jelena Karanović* and Aleksandra Nikolić

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Belgrade, Serbia
jelena.karanovic@imgge.bg.ac.rs

While the incidence of late-onset (≥50 years) colon cancer (LOCC) has been decreased worldwide, the number of early-onset (<50 years) colon cancer (EOCC) is increasing. Cluster of differentiation (CD) markers are widely and successfully used for identification of leukocyte population using flow cytometry in order to differentiate phenotypes among blood cancers and to address appropriate treatments. However, the alternations in their expression levels have been observed in non-blood cancers as well, including colon cancer, which enhance their diagnostic, prognostic and therapeutic biomarker potential.
The aim of this research was to identify potential CD biomarkers for EOCC and LOCC using open access to expression profiling by high throughput sequencing in order to conduct further experimental exploration.
Transcriptomic data of dataset GSE240623, obtained from formalin-fixed paraffin-embedded tumor tissue samples from 13 EOCC and 13 LOCC patients, and their paired-adjacent normal colon tissues, from Gene Expression Omnibus (GEO) database was used. Differentially expressed genes for EOCC and LOCC paired groups were obtained using DESeq2 package in R software with log2 fold change threshold set to 1. Up and downregulated genes were subsequently filtered only by CD markers for both comparisons. In EOCC patients, CD79B, CD22, CD19, CD79A, CD37 and CD48 were downregulated, while CD276 was upregulated in tumor tissue compared to control tissue. On the other hand, only CD33 was downregulated in LOCC tumor tissue compared to normal colon tissue. Finally, CD27 was downregulated in tumor tissues of both groups, EOCC and LOCC, in comparison to their counterparts. In this study, we have identified nine CD markers that have potential diagnostic, prognostic and/or therapeutic significance for colon cancer and will be further examined in *in silico* and *in vitro* study.

Keywords: colon cancer, biomarkers, GEO database, transcriptomic data

Acknowledgement: 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).

Structural characteristics of YtnP lactonase originating from *Stenotrophomonas maltophilia* 6960

Jovana Curcic[1,*], Milka Malesevic[1] and Branko Jovcic[2]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[2] Faculty of Biology, University of Belgrade
jcurcic@imgge.bg.ac.rs

Antibiotic resistance represents a serious global health threat. Available antibiotics progressively lose efficacy against many bacterial pathogens. This phenomenon underscores the urgent need for alternative strategies to combat bacterial infections. Progress in understanding the regulation of bacterial pathogenicity has prompted researchers to explore potential antivirulence drugs as a promising alternative. Quorum quenching enzymes capable of degrading *N*-acyl homoserine lactones can impede bacterial virulence and biofilm formation by disrupting cell-to-cell communication. In this study, we employed in silico structural characterization of YtnP lactonase originating from *Stenotrophomonas maltophilia* 6960, utilizing various online software tools, including AlphaFold2, I-TASSER, PSIPRED, Phyre2, and SWISS-MODEL algorithms. The results obtained from these programs were compared to each other.

The analysis revealed a 278 amino acid residue protein with a molecular weight of 31.02 kDa, predicted to be a transmembrane protein with an N-terminal extracellular domain and a C-terminal cytoplasmic domain, predominantly comprised of extracellular amino acid residues. Experimental validation demonstrated the quorum quenching activity of *S. maltophilia* 6960 towards exogenous AHLs, supporting the predicted role of YtnP lactonase in modulating quorum sensing of the surrounding bacteria. Furthermore, the observation of QQ activity in the crude extract and not in the cell-free supernatant of bacterial strain 6960 indicates that the YtnP lactonase is active within the bacterial cells. Secondary structure predictions revealed a balanced distribution of alpha-helices and beta-sheets, while tertiary structure predictions suggested a homodimeric configuration with four Zn^{2+} binding sites.

These findings, which combine in silico predictions with experimental validation, provide a solid foundation for further exploration in the development of effective antivirulence therapeutics. Leveraging *in silico* methodologies to predict and characterize the functional properties of potential antivirulence agents holds promise for accelerating the translation of research findings into clinical applications.

Keywords: bioinformatics, *in silico* prediction, quorum quenching, antivirulence therapeutics, lactonase

Acknowledgement: This study was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Agreement no - 451-03-47/2023-01/200042

Non-coding transcripts of protein-coding genes as novel biomarkers for colorectal cancer diagnosis

Jovana Despotović*, Sandra Dragičević, Tamara Babić, Dunja Pavlović, Jelena Karanović and Aleksandra Nikolić

Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
jovanadespotovic@imgge.bg.ac.rs

Recent research shows that non-coding RNA transcripts of protein-coding genes could be an emerging novel class of diagnostic biomarkers. This study aimed to identify the most promising biomarkers for colorectal cancer (CRC) screening among upregulated non-coding transcripts of protein-coding genes in malignant CRC cells in comparison to non-malignant cells grown in 3D. Malignant CRC cell lines HCT116, DLD-1 and SW620, and a non-malignant human colon epithelial cell line HCEC-1CT, were cultivated in 3D as spheroids in 24-well plates with low attachment surface for 7 days. RNA sequencing was performed on ribosomal-depleted total RNA using Illumina's NovaSeq6000 platform that generated paired-end 150bp reads. Highly upregulated transcripts (>10 FPKM) present in all malignant cell lines and absent in non-malignant cell line were filtered and analyzed by a set of in silico tools to filter the best candidates for further validation studies. The publicly available GSE164541 set consisting of triplicate tissue samples of normal, adenoma and primary CRC tissues collected from five patients with CRC was used for validation. As a result, 5 transcripts with retained introns ANXA3-204, LLGL2-207, KRT19-204, KRT18-206 and KRT8-213 were analyzed by in silico tools. Only ANXA3-204 was classified as non-coding according to both CPC2 and LGC online coding potential prediction tools. Nucleus was predicted as subcellular localization for ANXA3-204 by AnnoLnc2. Additionally, ANXA3-204 expression was upregulated in adenoma ($p=0.04$) and CRC tissue samples, although statistical significance was not reached for CRC, in comparison to normal tissue samples in the validation set. Transcriptomic analysis revealed that non-coding ANXA3-204 transcript was highly upregulated in malignant CRC cell lines and adenomas compared to control samples, making ANXA3-204 a potential candidate for early CRC screening. Further studies are needed to confirm diagnostic potential and regulatory role of ANXA3-204.

Keywords: non-coding RNA, colorectal cancer, biomarker, ANXA3, retained intron

Acknowledgement: This work was funded by the Ministry of Science, Technological Development, and Innovation of the Republic of Serbia (Contract No. 451-03-66/2024-03/200042), and by the Science Fund of the Republic of Serbia, PROMIS, #6052315, SENSOGENE.

Characterizing Somatic Mutation Clusters in Cancers Enriched with APOBEC Mutagenesis

Gennady V. Ponomarev[1], Fedor M. Kazanov[2] and Marat D. Kazanov[1,3,4,5,*]

[1] A.A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia
[2] "Foxford" High School, Moscow, Russia
[3] Skolkovo Institute of Science and Technology, Moscow, Russia
[4] Dmitry Rogachev National Medical Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia
[5] Sabanci University, Istanbul, Turkey
marat.kazanov@sabanciuniv.edu

The APOBEC family of cytidine deaminases plays an active role in the human immune system, combating viruses and transposable elements. These enzymes convert cytosine to uracil, which often leads to C->T or C->G mutations within the specific nucleotide contexts (TCN). Clusters of such mutations have been identified in the genomes of various cancer types, including bladder, breast, cervical, head and neck, and lung cancers. Previous studies have shown that these APOBEC-induced mutation clusters are not uniformly distributed across the genome. To analyze their distribution, we developed a method for detecting APOBEC-induced clusters in cancer genomes.

The method presented was developed using a subset of cancer samples from the PCAWG project, which exhibited a strong enrichment of APOBEC-signature mutations. Mutations were iteratively merged into clusters using a distance threshold determined through chromosome-specific simulations that matched the actual number of mutations. We constructed distributions of distances between neighboring mutations and derived distance thresholds for several significance levels (5%, 1%, 0.1%). We also separately estimated the heterogeneity of APOBEC mutagenesis along the genome, adjusting the distance threshold accordingly. This method was applied to the entire PCAWG dataset, identifying traces of APOBEC mutagenesis in additional cancer types.

Keywords: cancer bioinformatics, APOBEC, mutagenesis, mutation clusters.

Acknowledgement: This study was supported by Scientific and Technological Research Council of Turkey (TUBITAK) under the Grant Number 123E476. The authors thank to TUBITAK for their supports.

Machine learning methods for metabolite biomarkers detection

Miličić Lucija[1,*], Kovačević Jovana[1,2] and Kovačević Vladimir[2]

[1] Faculty of Mathematics, University of Belgrade, Belgrade, Serbia
[2] Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
lucija.milicic@matf.bg.ac.rs

Metabolites provide a unique view of the state of the entire organism. These small molecules produced by cellular processes can serve as indicators of a significant change in the body. The latest technological advances enabled measurement of up to a thousand metabolites from the blood, which paved the way for their usage as biomarkers or therapeutic activity indicators. The obtained metabolomic data requires special statistical and machine learning techniques for analyzing datasets with large number of features.
In our study we propose the methodology for processing metabolomics datasets with samples originating from groups with different phenotypes (e.g. disease and control group) and detecting metabolite candidates for potential biomarkers. We used preeclampsia datasets as a case study to test our methodology. The research focus was to determine whether any of the measured metabolites could indicate the onset of preeclampsia, and if so, to identify the most significant ones.
The approach to this problem involved developing a XGBoost classifier that would predict whether a patient has preeclampsia based on measured concentration of metabolites. For addressing the high dimensionality in the dataset, feature selector mRMR (minimum Redundancy - Maximum Relevance) was applied. The resulting model with an accuracy of 0.74 and ROC-AUC score of 0.8 on the test data, was used to identify the most important features that represent potential biomarker candidates. Statistical tests such as the T-test and Mann-Whitney test additionally confirmed a significant difference in the distribution of concentration of these metabolites between patients with and without preeclampsia. We detected increased concentration of specific fatty acids, along with cortisol, the stress hormone, in patients with preeclampsia. Further research will focus on understanding the mechanisms underlying these changes and their clinical relevance.

Keywords: bioinformatics, data mining, machine learning, metabolomics, preeclampsia

Acknowledgement: The authors want to thank The Chinese University of Hong Kong for collecting and publishing metabolomics data used in this research.

Molecular genetic basis of childhood epilepsy in Serbia: utility of clinical and whole exome sequencing

Andjelkovic M[1,*], Klaassen K[1], Skakic A[1], Marjanovic I[1], Kravljanac R[2,3], Djordjevic M[2,3], Vucetic Tadic B[2,3], Kecman B[2], Pavlovic S[1] and Stojiljkovic M[1]

[1] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia;
[2] Mother and Child Healthcare Institute „dr Vukan Cupic", School of Medicine, University of Belgrade, Serbia;
[3] Faculty of Medicine, University of Belgrade, Belgrade, Serbia
marina.andjelkovic@imgge.bg.ac.rs

Childhood epilepsies are caused by heterogeneous underlying disorders where approximately 40% of the origins of epilepsy can be attributed to genetic factors. The application of next-generation sequencing (NGS) has revolutionized molecular diagnostics and has enabled identification of disease-causing genes and variants in epilepsies.
In our study, 55 children with epilepsy of unknown etiology were analyzed combining clinical-exome (CES) and whole-exome sequencing (WES). Novel variants were characterized using various *in silico* algorithms for pathogenicity and structure prediction. Molecular genetic cause of epilepsy was identified in 28 patients and the overall diagnostic success rate was 50.9%. We identified variants in 22 different genes associated with epilepsy that correlate well with the described phenotype. *SCN1A* gene variants were found in 5 unrelated patients, while *ALDH7A1* and *KCNQ2* gene variants were found twice. In the other 19 genes, variants were found only in a single patient. This includes genes: *ASH1L, CSNK2B RHOBTB2* and *SLC13A5*, which have only recently been associated with epilepsy. Almost half of diagnosed patients (46.4%) carried novel variants. Interestingly, identification of variants in *ALDH7A1, KCNQ2, PNPO, SCN1A* and *SCN2A* gene directed therapy decision of 11 children from our study, including four children who all carry novel *SCN1A* genetic variants.
Our study emphasizes the importance of NGS in diagnosing childhood epilepsy. With an increasing number of genes associated with epilepsy, comprehensive analysis using CES and WES is crucial for high diagnostic success. Given the expansion of molecular-based approaches, each newly identified genetic variant could become a potential therapeutic target.

Keywords: childhood epilepsy, monogenic disease, CES, WES, novel genetic variants

Acknowledgement: This work was supported by Ministry of Science, Technological Development and Innovations Republic of Serbia [Number: 451-03-47/2023-01/200042].

Transcriptome-wide detection of RNA cleavage sites revealed tRNA cleavage by target-activated CRISPR-Cas13a effector

Matvei Kolesnik[1,*], Ishita Jain[2], Ekaterina Semenova[2] and Konstantin Severinov[2]

[1] Center for Molecular and Cellular Biology, Skolkovo Institute of Science and Technology, Moscow, Russia
[2] Waksman Institute, Rutgers, The State University of New Jersey, NJ 08854 USA
matvei.kolesnik@skoltech.ru

Type VI CRISPR-Cas systems exclusively recognize and cleave RNA molecules. A distinct feature of Type VI systems is collateral RNA damage. Specifically, the binding of a target transcript by Cas13a, charged with cognate CRISPR RNA (crRNA), activates the Cas13a enzyme, turning it into an active ribonuclease that mediates the cleavage of non-complementary RNA molecules. Previously, Cas13a-mediated collateral RNA cleavage was observed in in vitro experiments and was described to be nonspecific. In *Escherichia coli*, targeting of nonessential transcripts by heterologously expressed *Leptotrichia shahii* Cas13a enzyme (LshCas13a) leads to cell growth retardation, which was proposed to be a consequence of collateral degradation of essential cellular transcripts. However, the direct link between collateral RNA cleavage and cell growth retardation was not established. Specifically, the products of collateral RNA cleavage mediated by target-activated Cas13a enzyme were not identified in living cells.

To detect RNA cleavage sites associated with collateral Cas13a activity, a specific approach based on high-throughput RNA sequencing was developed. This approach was successfully applied to detect RNA cleavage sites introduced by target-activated LshCas13a enzyme in both in vivo and in vitro experiments. In *E. coli* cells, the target-activated LshCas13a enzyme cleaves tRNA molecules within anticodon loops, leading to protein synthesis inhibition and slowing down cell growth. Additionally, LshCas13a-mediated collateral tRNA cleavage indirectly activates cellular ribonucleases encoded by Type II toxin-antitoxin systems.

Together, the results suggest that the *L. shahii* Type VI CRISPR-Cas system mediates the immune response by inhibiting translation through collateral tRNA cleavage.

Keywords: CRISPR-Cas systems, RNA-Seq, Cas13

Acknowledgement: I want to thank Prof. Pavel Mazin for his comments on the RNA-Seq data analysis.

Transcriptome profiling of pharmacological manipulation of zebrafish tailfin regeneration

Mila Ljujić*, Jelena Kušić Tišma, Bojan Ilić and Aleksandra Divac Rankov

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Belgrade, Serbia
milaljujic@imgge.bg.ac.rs

Zebrafish (*Danio rerio*) have a remarkable regenerative capacity of various tissues and organs, making them a widely used model organism for studying tissue regeneration and therapeutic development through pharmacological screens. The zebrafish larval fin consists of a two-layered, infolded epithelium and its regeneration after amputation is completed within just 3 days. We utilized this model to study the effect of human protein alpha-1 antitrypsin (AAT) on wound healing. This study aimed to identify the effect of AAT on transcriptional profiles during zebrafish larvae fin regeneration at 24h post amputation. Caudal fin was amputated at 48h post fertilization and larvae were treated with 2mg/ml AAT for 24h.

Pools of 24 larvae were collected at 24h post amputation and total RNAs were extracted using TRIzol Reagent. RNAseq was performed using Illumina Novaseq6000. Adapter trimming and low-quality reads were removed from raw data with fastp and reads were aligned to Danio rerio genome assembly GRCz11 by STAR aligner. Counts of differentially expressed transcripts were calculated at gene level with FeatureCount. Differential expression was analysed by DESeq2 package after correction for batch effect by ComBat-seq from sva bioconductor package. The threshold for significantly differential expression was set as the adjusted p-value < 0.01.

At 24hpa there were 185 differentially expressed genes (89 upregulated and 96 downregulated) in treated larvae compared to the untreated. *Mgst2*, *Apobb.2*, and *Prss59.2* were in the top 10 downregulated genes while *Col9a1b*, *Col8a1a*, and *Matn3a* were in the top 10 upregulated genes. Enrichment analysis by Enrichr showed that key gene onthologies in AAT treated zebrafish larvae were intermediate filament, collagen containing extracellular matrix, cellular response to glucocorticoid and endoplasmic reticulum lumen among others. Our data support further evaluation of AAT as a potential promoter of wound healing.

Keywords: RNA-sequencing, zebrafish, regeneration, wound healing

Acknowledgement: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-66/2024-03/200042)

Exploration of Intrinsic Disorder Regions through Classification of Intrinsically Disordered Proteins Using PPI Network Structure and Sequence Attributes: A Case Study

Milana Grbić*, Milan Predojević, Nenad Vilendečić and Dragan Matić

Faculty of Natural Science and Mathematics, University of Banja Luka, Banja Luka, Bosnia and Herzegovina
milana.grbic@pmf.unibl.org

In this study, the prediction of Intrinsically Disordered Proteins (IDPs) was explored by utilizing the structure of Protein-Protein Interaction (PPI) networks and sequence characteristics. A weighted PPI network, where edge weights represented gene co-expression information between two proteins, was used to extract attributes related to protein topological properties via the node2vec+ tool. Additionally, attributes derived from primary sequence information were incorporated, focusing on amino acid properties such as order/disorder promotion (Type A attributes) and physicochemical properties including aromatic/aliphatic, polar/non-polar, non-zero/zero, hydrophobic/hydrophilic, and positive/negative (Type B attributes).
Proteins were classified into IDP and non-IDP categories using a K-Nearest Neighbors (KNN) classifier under four scenarios: (1) based solely on network attributes, (2) incorporating network attributes and sequence Type A attributes, (3) incorporating network attributes and sequence Type B attributes, and (4) considering network attributes along with both sequence Type A and Type B attributes. Proteins misclassified as IDPs in these scenarios were further examined using the IUPred2 tool, which revealed that only a subset of these proteins indeed possessed intrinsic disorder regions (IDRs) along their sequences.
This study was conducted as a case study using the PPI network model of the yeast organism from the BioGRID database, with the list of yeast IDP proteins sourced from the DisProt database. Gene co-expression information was obtained using the SPELL tool.

Keywords: Intrinsically Disordered Proteins (IDPs), Protein-Protein Interaction (PPI) Networks, Sequence Attributes, IDP prediction

Acknowledgement: This research is supported by a project titled "Support to COST action: Information, Coding and Biological Function: The Dynamics of Life" funded by the Ministry of Scientific and Technological Development and Higher Education, Government of Republic of Srpska, B&H. Also, it is supported by two projects funded by Ministry of Civil Affairs, B&H: "Support for the Participation of the Research Team of Bosnia and Herzegovina in COST Action CA21169: Information, Coding, and Biological Function: the Dynamics of Life", and "Support for the Implementation of COST Action CA21160 Non-globular Proteins in the Era of Machine Learning in Bosnia and Herzegovina (ML4NGPB&H)".

Navigating ELSI for FAIR multiomics data management within STEPUPIORS international rectal cancer project

Miljana Tanić[1], Mariska Bierkens[2], Marko Radulović[1], Aleksandra Stanojević[1], Mladen Marinković[3], Ana Krivokuća[1], Radmila Janković[1], Ana Đurić[1], Sergi Castellvi-Bel[4], Jerome Zoidakis[5,6], Remond J. Fijneman[2] and Milena Čavić[1]

[1] Experimental Oncology Department, Institute for Oncology and Radiology of Serbia (IORS), Belgrade, Serbia
[2] Department of Pathology, Netherlands Cancer Institute (NKI), Amsterdam, Netherlands
[3] Department of Radiation Oncology, Institute for Oncology and Radiology of Serbia, Belgrade, Serbia
[4] Gastroenterology Department, Fundació Recerca Clínic Barcelona-Institut d'Investigacions Biomèdiques August Pi i Sunyer (FRCB-IDIBAPS), Barcelona, Spain
[5] Department of Biotechnology, Biomedical Research Foundation, Academy of Athens (BRFAA), Athens, Greece.
[6] Department of Biology, National and Kapodistrian University of Athens, Athens, Greece
tanic.miljana@ncrc.ac.rs

The STEPUPIORS project entails a multi-omics approach to studying neoadjuvant chemoradiotherapy (nCRT) response in locally-advanced rectal cancer (LARC) patients through retrospective and prospective longitudinal sample and data collection. Through a multinational collaboration including four partnering institutions (IORS - Serbia, NKI - Netherlands, BRFAA - Greece, FRCB-IDIBAPS - Spain) different layers of omics data are being generated and integrated to derive biomarkers and models for prediction of patient's response to nCRT. Consideration of Ethical, legal and societal issues (ELSI) is mandatory for projects dealing with personal and sensitive data, and collection of biological material from patients. The international nature of the project necessitates compliance with both national (Serbian) and EU laws and regulatory provisions governing data privacy and patient protection.

Ethics approvals were sought with the IORS' local Committee for retrospective use of samples and health records, and for prospective collection and biobanking of samples and associated metadata, supplying patient information sheet and 2-tiered informed consent form. Prospectively collected samples and data will be managed by a Laboratory Information Management System (LIMS) acquired within the newly established IORS Biobank. International collaboration was facilitated through signing of Consortium and Joint controller agreements specifying terms for transfer of material including data, non-disclosure of information and data processing activities.

A Data Management Plan (DMP) was made considering a wide range of data types including structured information from patients' medical records, radiology images, genomics, transcriptomics and proteomics data. DMP contained detailed descriptions of type of processing, source of data, data format and quantity. Means of maintaining confidentiality were described, which include access control, pseudonymization, separate processing of data collected for different purposes. Free access to data was achieved through deposition of raw data in appropriate repositories (Zenodo, EGA, cBioPortal) and as Supplementary data

in open access publications in accordance with EU's Open Science policy and in line with Findable, Accessible, Interoperable Research data (FAIR) principles.
Here we provide the roadmap and examples for navigating the complex ELSI landscape required for FAIR biomedical and multi-omics data management in accordance with the applicable regulatory provisions relating to the protection of the personal data and to medical confidentiality.

Keywords: data management, regulatory, ethics, privacy, omics

Acknowledgement: Horizon Europe Project STEPUPIORS (101079217)

Enhancing Cancer Genomics: A Pipeline for Spatial Transcriptomics Analysis on the CGC

Miona Rankovic, Nevena Vukojicic*, Nevena Ilic Raicevic, Vida Matovic and Ana Mijalkovic Lazic

Velsera, Belgrade, Serbia
nevena.vukojicic@velsera.com

Spatial transcriptomics field has grown significantly in recent years. This hybrid method, inspired by in situ hybridization and next-generation sequencing, particularly single-cell RNA sequencing (scRNA-seq), enables whole transcriptome profiling while maintaining spatial context at high resolutions, offering new insights in cancer research.

We present a highly configurable sequencing-based technology solution for comprehensive spatial analysis. Available on the NCI-funded Cancer Genomics Cloud (CGC) platform by Seven Bridges, this pipeline provides a collaborative cloud infrastructure. The CGC platform integrates computation, over 1000 bioinformatics workflows, and 4+ PB of data, making Cancer Research Data Commons (CRDC) datasets accessible from any environment.

Developed with widely adopted packages, this pipeline processes datasets from leading technologies, 10x and Slide-seq. It includes steps such as quality control, data preprocessing, dimensionality reduction, cluster identification, detection of spatially variable features, and integration with scRNA-seq references. The pipeline is highly configurable, allowing various settings to be optimized for better results, and some specific components can be selectively executed. Key steps are visually represented for detailed insights.

Here, we demonstrate spatial transcriptomics analysis flow on publicly available datasets using this pipeline, showing the impact of different settings on analysis outcomes. We identify spatially variable genes with distinct tissue localization and integrate data to predict cell type composition within spatial domains.

Spatial transcriptomics analysis significantly enhances cancer research by characterizing tumor microenvironments, discovering novel biomarkers, and clarifying drug resistance mechanisms. This CGC-hosted workflow is expected to contribute to significant advancements in understanding complex spatial relationships within tissues.

Keywords: spatial transcriptomics, single-cell omics, cloud computing

Impact of 3D chromatin structure on cancer mutation patterns and tissue-of-origin prediction

Paula Štancl* and Rosa Karlić

Faculty of Science, University of Zagreb, Croatia
pstancl@bioinfo.hr

Tissue-of-origin (TOO) detection poses a challenge for effective therapy selection due to the heterogeneity and distinct genomic and molecular characteristics of carcinoma of unknown primary site (CUP). Machine-learning algorithms utilizing mutational landscape data from whole-genome sequencing and normal tissue epigenetic features have shown promise in predicting TOO. These models leverage the association between mutation densities, regional histone modifications, and the non-uniform distribution of mutations across 1 MB genomic regions and tumor types. However, some cancer types remain difficult to classify accurately. To address this, we developed TOO models utilizing tissue-specific topologically associated domains (TADs). Since TADs represent fundamental units of 3D genome architecture, we investigated whether the TOO prediction can be improved by using TADs (TAD model) or genes clustered based on their location in TADs (TAD gene model).
We analyzed publicly available liver cancer cohorts from the International Cancer Genome Consortium, obtaining ChIP-seq data for six histone modifications and input controls from the Roadmap Epigenomics project. Tissue-specific TADs were downloaded from TADBK and the 3D Genome Browser. We used a multiple linear regression model with 10-fold cross-validation to compute the amount of variance of aggregated mutations across various TADs or TAD-based gene clusters explained by the epigenome of each normal tissue. The model with the highest variance represents the TOO for a specific cancer type. The results demonstrated consistent correct TOO prediction across all tissue-specific TADs identified using various tools. Although the overall accuracy of TAD models did not significantly differ from the original model developed using 1 MB region-based predictions, TAD gene models showed a significant increase in correct TOO prediction compared to other gene models we developed. Genes located in TADs where the number of predicted mutations was lower than the observed number were associated with cancer development and progression, indicating that this type of analysis can facilitate the identification of structural units that influence carcinogenesis.
Overall, the results show that we can use the cell's epigenome and cancer's mutation profile based on TADs to predict the tissue-of-origin and use the developed models to analyze the mechanisms of cancer initiation and progression.

Keywords: tissue-of-origin, machine learning, topologically-associated domains, mutational landscape, epigenome

Unsupervised domain adaptation methods for cross-species transfer of regulatory code signals

Pavel Latyshev[1,*], Fedor Pavlov[1,*], Alan Herbert[1,2] and Maria Poptsova[1,*]

[1] Laboratory of Bioinformatics, Faculty of Computer Science, HSE University, Moscow, Russia
[2] InsideOutBio, Charlestown, MA, United States
platyshev@hse.ru

Due to advances in NGS technologies whole-genome maps of various functional genomic elements were generated for a dozen of species, however experiments are still expensive and are not available for many species of interest. Deep learning methods became the state-of-the-art computational methods to analyze the available data, but the focus is often only on the species studied. Here we take advantage of the progresses in Transfer Learning in the area of Unsupervised Domain Adaption (UDA) and tested nine UDA methods for prediction of regulatory code signals for genomes of other species. We tested each deep learning implementation by training the model on experimental data from one species, then refined the model using the genome sequence of the target species for which we wanted to make predictions. Among nine tested domain adaptation architectures non-adversarial methods Minimum Class Confusion (MCC) and Deep Adaptation Network (DAN) significantly outperformed others. Conditional Domain Adversarial Network (CDAN) appeared as the third best architecture. Here we provide an empirical assessment of each approach using real world data. The different approaches were tested on ChIP-seq data for transcription factor binding sites and histone marks on human and mouse genomes, but is generalizable to any cross-species transfer of interest. We tested the efficiency of each method using species where experimental data was available for both. The results allows us to assess how well each implementation will work for species for which only limited experimental data is available and will inform the design of future experiments in these understudied organisms. Overall, our results proved the validity of UDA methods for generation of missing experimental data for histone marks and transcription factor binding sites in various genomes and highlights how robust the various approaches are to data that is incomplete, noisy and susceptible to analytic bias.

Keywords: bioinformatics, transfer learning, domain adaptation, histone marks, transcription factors

De novo genome sequencing for endangered bird of prey species

Erić Pavle[1,*], Marija Tanasković[1], Aleksandra Patenković[1], Katarina Erić[1], Irena Hribšek[1], Kristijan Ovari[2] and Slobodan Davidović[1]

[1] Department of Genetics of Populations and Ecogenotoxicology, Institute for Biological Research "Siniša Stanković" - National Institute of the Republic of Serbia, University of Belgrade, Belgrade, Serbia
[2] Belgrade ZOO, Belgrade, Serbia
pavle.eric@ibiss.bg.ac.rs

The Eastern Imperial Eagle (*Aquila heliaca*) is a large migratory bird of prey, with breeding sites spanning from eastern Czechia and Austria to Northwestern China and Mongolia. Due to the decline of its populations throughout its area of distribution, the IUCN Red List categorized species as vulnerable. Thus, it has become a subject of numerous conservation efforts. To develop effective conservation strategies, it is crucial to have a comprehensive understanding of the genetic variability of these populations. Establishing a reference genome serves as a cornerstone for conservationists, offering a starting tool to assess population dynamics, adaptive potential and evolutionary history with further analyses. Therefore, we performed the whole genome sequencing of *A. heliaca*.
The genome sequencing of a male *A. heliaca* was conducted using Illumina paired-end 150bp short reads, and *de novo* assembly was conducted as there is no reference genome available. After the use of paired-end information for scaffolding the assembly remained very fragmented, with the genome being represented with hundreds of thousands of contigs, primarily due to the inherent limitations of short-read sequencing in resolving repetitive regions and regions with strong nucleotide composition bias. To enhance scaffolding, we used a chromosome-level assembly of a closely related species, *Aquila chrysaetos*, available in the GenBank. BLAST analysis revealed high sequence similarity (~94%) between sequences from our assembly compared to the *A. chrysaetos* reference genome. The absence of major rearrangements or inversions in the selected contigs supported the usage of the *A. chrysaetos* reference genome for scaffolding. Thus, we generated a chromosome-level assembly for *A. heliaca*, encompassing 26 autosomes and the Z sex chromosome. Even though our assembly contained a few thousand unplaced scaffolds ranging in size from over 700Kb to very small fragments, the vast majority (>98%) of the assembly was assigned to 27 chromosomal-level scaffolds. The assembly demonstrated a completeness score of 97.2% according to the BUSCO assessment.

Keywords: *Aquila heliaca*, WGS, De-novo genome assembly, conservation

Acknowledgement: We would like to express our sincere gratitude to Professor Goran Rakočević for his invaluable help and suggestions in tackling the assembly process. We are also grateful to the RAF School of Computing for granting us access to their server. Their support was pivotal in enabling the assembly of the genome, and we greatly appreciate their contributions.

Mouse Tissue of Origin Single Cell Classification System

Sen Lin[1], Vladimir Brusic[2] and Tianyi Qiu[3]

[1] School of Computer Science, University of Nottingham Ningbo, Ningbo, China
[2] School of Economics, University of Nottingham Ningbo, Ningbo, China
[3] Institute of Clinical Science, Zhongshan Hospital,
Fudan University, Shanghai, China
SEN.LIN2@nottingham.edu.cn
vladimir.brusic@nottingham.edu.cn

Single cell transcriptomics (scRNA-seq) technology can concurrently measure gene expression from hundreds of thousands of individual cells. The aim of this project is to build a system for classification of tissue and organ of origin of single cell types and subtypes for the mouse samples. The classification system maps the cells using scRNA-seq data and supervised machine learning methods.
The first step was to develop a hierarchical mouse cell type map to comprehend the biology and heterogeneity of different cell types. The second step was to build mouse scRNA-seq reference datasets containing high-quality and well-annotated scRNA-seq datasets representing different mouse strains, animal ages, sexes, and biological conditions. Third, his work established computational workflows that integrated standardization, quality control, clustering, annotation, and classification model building of scRNA-seq data. The standardization work involved a protocol for data processing that mapped different gene versions, names, quantities, and data formats to standardized formats. Quality control involved data visualization, filtering errors and uninformative measurements based on cell distribution, detection of outliers, and using standard gene markers to filter unrelated cells. The final step was to build a classification system based on supervised machine learning. The step deployed the feed-forward multilayer artificial neural network with logistic regression learning algorithms. The resulting model built using approximately 117 thousand cells from 15 different tissues achieved classification accuracy of 93.6%. Most misclassified cells were immune cells that are known to migrate across tissues and organs. The classification accuracy of tissue of origin of immune cells differs according to the tissue with high classification accuracy (>95%) for aorta, heart, islets, liver, pancreas, and thymus. Classification accuracy of immune cells from blood, colon, lung, pancreatic lymph nodes, small intestine, and spleen as tissue of origin was lower.
This research demonstrated that supervised machine learning methods could achieve high accuracy in classifying mouse single cell types from most tissues and organs.

Keywords: bioinformatics, data mining, computer science, single cell transcriptome, machine learning

Acknowledgement: This work was supported by the University of Nottingham Ningbo High-Flyer PhD Scholarship.

LEA4 protein group member from resurrection plant *Ramonda serbica* Panč. – production and *in silico* characterization

Tatiana Ilina*, Ana Pantelić, Dejana Milić and Marija Vidović

Institute of Molecular Genetics and Genetic Engineering,
University of Belgrade, Belgrade, Serbia
tanyailina134@gmail.com

Ramonda serbica Panč., an ancient resurrection plant, can withstand extended periods of desiccation and reestablish metabolic activity shortly upon watering. One of the key players of desiccation resistance in resurrection species is the accumulation of protective late embryogenesis abundant proteins (LEAPs). During cellular dehydration, these intrinsically disordered proteins (IDPs) may stabilize the native structures of proteins and membranes. Differential transcriptome and proteome analyses revealed that members of the LEA4 protein family represent the majority of desiccation-inducible LEAPs.
The aim of this work was to *in silico* characterize and recombinantly produce a member of the LEA4 protein family group – LEA_301. This protein is predicted to be highly disordered (above 85%). Its sequence is composed of predominantly charged and polar amino acids (39% of the sequence is charged, 12% of the sequence is lysine), with acidic pI (5.92). According to secondary structure predictors (JPred, PsiPred, Phyer2, Sopma, and FELLS), LEAP_301 exhibits high propensity to form amphipathic α-helix (above 95 %). This can be important for their function during desiccation, when this protein may undergo a disorder-to-order transition, and stabilize biomolecules. To experimentally validate the secondary structure of this protein and assume its physiological role, we produced high yield and purity of LEA_301 by recombinant DNA technology.
Bacterial cells (*Escherichia coli* BL21(DE3)) were transformed with the vector with the LEA_301-6xHis gene, and optimization of protein expression was done. Final purification was done by immobilized metal ion affinity chromatography. The sequence was validated by Western blot and mass spectrometry. Obtained LEA_301 will be structurally characterized using methods for secondary structure determination such as Fourier-transform infrared spectroscopy and circular dichroism spectroscopy and compared with the *in silico* results.

Keywords: bioinformatics, LEA proteins, secondary structure prediction, intrinsically disordered proteins, production of recombinant proteins

Acknowledgement: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-66/2024-03/200042) and by the Science Fund of the Republic of Serbia-RS (PROMIS project LEAPSyn-SCI, grant no. 6039663).

SUPPORTED BY

Ministry of Science, Technological Development and Innovation Republic of Serbia

THE WORLD BANK IBRD • IDA

#ЕУ ЗА ТЕБЕ

CHAMBER OF COMMERCE AND INDUSTRY OF SERBIA

GOLD SPONSORS

BGI MGI

ELTA'90MS More than Technology

SILVER SPONSORS

PacBio

East Diagnostics

ALFA GENETICS

BRONZE SPONSORS

vivogen

LKB

Telekom Srbija

Altium

SPONSORS

Labena

Alfamed

GALEN - FOKUS

KEFO

SUPERLAB® Your lab – Our passion

РНИДС

RTC

BIOMEDICA

ISBN: 978-86-82679-16-5