Symposium of Polish Bioinformatics Society
September 11-13, 2024, Warsaw

Book of Abstracts

Faculty of Electronics and Information Technology
WARSAW UNIVERSITY OF TECHNOLOGY

bioforum Central European

aws

AstraZeneca

https://www.ptbi.org.pl/website/conferences/ptbi24

Symposium Day 1-11th September, 2024					
Wednesday 11.09.2024					
Hours	**Session**	**Presentation Tittle**	**Speaker**	**Affilliation**	**Chairperson**
09:00-10:45	**Registration**				
10:45-11:00	**Welcome**				
11:00-11:50	**Honorary member**	**Coarse-Grained protein modeling**	Andrzej Koliński	University of Warsaw	Tomasz Gambin
11:50-12:10	**Session 1**	Understanding Protein Aggregation Through a Simple Coarse-Grained Model	Mohammad Saqib	University of Warsaw	
12:10-12:30		Modeling of RNA 3D structure and interactions, with restraints derived from experimental data	Janusz Bujnicki	International Institute of Molecular and Cell Biology	
12:30-12:50		SimRNA-sol: A novel method for simulating RNA folding with water and ions, and predicting tightly-bound solvent molecules in RNA 3D structures	Masoud Amiri Farsani	International Institute of Molecular and Cell Biology	
12:50-13:10	**Coffee break**				
13:10-13:30	**Session 2**	Structural and functional diversity across large protein databases	Paweł Szczerbiak	Sano Centre for Computational Medicine, Jagiellonian University	Paweł Łabaj
13:30-13:50		Strain-resolved metagenomics in analysis of gut microbiota dynamics within families	Katarzyna Sidorczuk	Norwich Research Park, UK	
13:50-14:10		MultiEM: Modelling Chromatin Structure from Nucleosomes to Chromosomal Territories	Sebastian Korsak	Warsaw University of Technology, University of Warsaw	
14:10-14:30		Computational modeling of chromatin three-dimensional structure from super-resolution microscopy	Zofia Parteka-Tojek	Warsaw University of Technology, University of Warsaw	
14:30-14:50		Microbiome Diversity and Intra-Community Synergies Across Climate-Defined Ecological Niches	Dagmara Błaszczyk	Jagiellonian University	
14:50-15:50	**Lunch break**				
15:50-16:20	**Sponsor**	**Generative AI in Health**	**Tomasz Żemojtel**	**Amazon Web Services**	Bartosz Wilczyński
16:30-18:30	**Poster session**				
18:30 - 00.00	**Bars/restaurant/pubs**				

Symposium Day 2-12th September, 2024					
Thursday 12.09.2024					
Hours	**Session**	**Presentation Tittle**	**Speaker**	**Affilliation**	**Chairperson**
9:00-9:50	**Keynote speaker**	**Fantastic world of tandem repeats and how to characterize them**	**Fritz Sedlazeck**	**Baylor College of Medicine**	Marta Kasprzak
9:50-10:10	**Session 3**	Quantitative analysis of tRNA modifications by nanopore RNA sequencing	Małgorzata Adamczyk	Warsaw University of Technology	
10:10-10:30		A comprehensive pipeline for gene expression and mutation profiling via targeted sequencing with unique molecular barcodes	Michał Marczyk	Silesian University of Technology	
10:30-10:50		Enhancing RNA-seq Fusion Detection: A Meta-Analysis of Benchmark Variability and Tool Performance	Iga Ostrowska	Warsaw University of Technology	
10:50-11:10	**Coffee break**				
11:10:11:30	**Session 4**	A three-level modeling for identifying important predictor variables in genome-wide association studies suffering from p >> n	Jakub Liu	Wroclaw University of Environmental and Life Sciences	Aleksandra Gruca
11:30-11:50		Efficient and Accurate LC-MS Feature Alignment Using Sliced-Wasserstein Distance.	Justyna Król	University of Warsaw	
11:50-12:10		Unsupervised learning for detecting relevant importance of pathway activity in individual cells based on scRNA-Seq data	Joanna Żyła	Silesian University of Technology	
12:10-12:30		Deep into the Dark Proteome: structural disorder analysis of low-complexity subtypes in protein sequences using the AlphaFold pLDDT metric	Barbara Ilnicka	Silesian University of Technology	
12:30-12:50		An automated analysis of homocoupling defects using MALDI-MS and open-source computer software	Maria Bochenek	University of Warsaw	
12:50-13:00	**Group photo**				
13:00-13:50	**Lunch break**				
13:50-14:40	**Invited talk**	**Omics Data Science**	**Catherine Suski-Grabowski**	University of Warsaw	Witold Rudnicki
14:40-15:10	**Sponsor**	**Leveraging genomics at scale for Drug Discovery: the AstraZeneca Genomics Initiative**	**Sebastian Wasilewski**	**AstraZeneca**	
15:10-15:30	**Coffee break**				
15:30-17:00	General Assembly of Polish BIoinformatics Society				
17:00 - 19:30	Social event, Warsaw Old City				
19:30-22:00	**Gala Dinner**				

Symposium Day 3-13th September, 2024					
Friday 13.09.2024					
Hours	**Session**	**Presentation Tittle**	**Speaker**	**Affilliation**	**Chairperson**
9:00-9:15	**Laureates**	Computational methods for anti-cancer drug sensitivity prediction	Krzysztof Koras	University of Warsaw	Aleksandra Świercz
9:15-9:30		Modeling and simulation of multi-agent systems representing processes in the RNA World hypothesis	Jarosław Synak	Poznan University of Technology	
9:30-9:40		Robinson-Foulds distance between phylogenetic networks and gene trees	Natalia Rutecka	University of Warsaw	
9:40-9:50		Algorithm for constructing a variation graph from a colored de Bruijn graph	Adam Cicherski	University of Warsaw	
9:50-10:00		Determining gene expression profile in pituitary somatotroph tumors and identification of their molecular subtypes	Julia Rymuza	University of Warsaw	
10:00-10:10		Creation and evaluation of an amino acid substitution matrix for low complexity fragments of proteins combined with the improvement of a method to compare these fragments	Maciej Dzikowski	University of Warsaw	
10:10-10:20		Evaluation of the Efficiency of Signaling Pathway Activation Scores Matrices Using Unsupervised Machine Learning Techniques in scRNA-Seq Data	Kamila Szumala	Silesian University of Technology	
10:20-10:30		Quality Assesment of 3D RNA Structures Using Graph Neural Networks	Bartosz Adamczyk	Poznan University of Technology	
10:30-10:45	**Coffee break**				
10:45-11:30	**Keynote speaker**	**Modeling spatial omics profiles of tumor microenvironments**	**Ewa Szczurek**	**University of Warsaw and Helmholtz Munich**	Robert Nowak
11:30-11:45	**Awards ceremony**				
11:45-12:00	**Closing remarks**				
12:00-13:00	**Lunch break**				

Contents

Day 1 - 11 September 2024

Keynote speaker - Honorary PTBI Member

Coarse-Grained protein modeling

Andrzej Koliński[1]

[1] *Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw*

The traditional computational modeling of protein structure, dynamics, and interactions remains difficult for many protein systems. This is mostly due to the size of protein conformational spaces and required simulation timescales that are still too large to be studied in atomistic detail. Lowering the level of protein representation and/or simplification of the force field opens up new possibilities for studying protein systems. Coarse-grained models (1) are computationally more effective and enable simulations of much longer time scales and/or larger sizes of the systems studied. Moreover, well-designed coarse-grained models of a not-too-low resolution enable reasonable reconstruction of modeled structures to all-atom resolution. This way efficient multiscale simulations of important biomolecular processes become possible. I will focus on the CABS model designed 20 years ago (2) and consistently improved in my laboratory. The CABS modeling tools use discrete representations of conformation space, statistical (knowledge-based) potentials, and Monte Carlo dynamics sampling schemes, applied to the single or multiple replicas of modeled objects. CABS model is unique and its flaws and advantages are discussed. The model can be efficiently used for simulations of protein folding (3), molecular docking (4), and large-scale dynamics (5). Finally, possible integrations of the CABS method with AI tools are briefly indicated.

(1) S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. Dawid & A. Kolinski, "Coarse-grained protein models and their applications", Chemical Reviews 116(14) 7898–7936 (2016)
(2) A. Kolinski, "Protein modeling and structure prediction with a reduced representation", Acta Biochimica Polonica 51:349-371 (2004)
(3) M. Blaszczyk, M. Jamroz, S. Kmiecik & A. Kolinski, "CABS-fold: server for de novo and consensus-based prediction of protein structure" Nucleic Acids Research 41(W1):W406-W411 (2013)
(4) M. Kurcinski, M. Jamroz, M. Blaszczyk, A. Kolinski & S. Kmiecik, "CABS-dock: web server for flexible docking of peptides to proteins without prior knowledge of the binding site", Nucleic Acids Research 43(W1): W419-W424 (2015)
(5) A. Kuriata, A. Gierut, T. Oleniecki, M. Ciemny, A. Kolinski, M. Kurcinski, S. Kmiecik, "CABS-flex 2.0: a web server for fast simulations of flexibility of protein structures", Nucleic Acids Research, 46(W1):W338-W343 (2018)

Session 1

Understanding Protein Aggregation Through a Simple Coarse-Grained Model

Mohammad Saqib[1] Dominik Gront[2]

[1] *Interdisciplinary Doctoral School, University of Warsaw*

[2] *Faculty of Chemistry, University of Warsaw*

Simple coarse-grained (CG) models provide valuable insights into the behavior and dynamics of proteins. In this study, we devised a straightforward CA-only model to simulate protein dynamics, focusing exclusively on C-alpha atoms. This approach streamlines the complexity of atomic-level simulations, making it easier to study protein behavior in various conditions. Even though our model is simplified, it uses structured algorithms and move proposals, such as TailMove and the newly added HingeMove, to simulate how protein ends and segments move inside proteins. Non-bonded interaction kernels, like excluded volume and C-alpha contact energy, are used to figure out the system's energy. These kernels measure potential energy based on the distances and repelling forces between atoms.

We conducted simulations at different protein concentrations to observe their effects on protein behavior. Results suggest that variations in concentration significantly impact protein dynamics and aggregation processes. Key observables, such as the center of mass, end-to-end distances, and radius of gyration, were tracked to provide a comprehensive understanding of the system. By employing Monte Carlo methods, our model explores various protein conformations, with moves accepted or rejected based on changes in the system's energy. Periodic boundary conditions and dedicated modules for rotations and translations ensure system stability throughout the simulation.

We hope that these simulations will enhance our understanding of protein aggregation, offering a simple yet effective tool for studying complex biophysical phenomena.

Keywords: Coarse-Grained Model, Protein Dynamics, C-alpha Atoms, Monte Carlo Methods, Protein Aggregation

Modeling of RNA 3D structure and interactions, with restraints derived from experimental data

Janusz M. Bujnicki[1]

[1] *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw, Poland. Email: janusz@iimcb.gov.pl*

Ribonucleic acid (RNA) molecules are master regulators of cells. They play key roles in many molecular processes: transmitting genetic information, sensing cellular signals, relaying responses, and even catalyzing chemical reactions. The function of RNA, especially its ability to interact with other molecules, is encoded in its sequence. To understand how these molecules carry out their biological tasks, we need detailed knowledge of RNA structure, dynamics, and thermodynamics. The latter largely determines how RNA folds and interacts within the cellular environment.

Experimentally determining these properties is challenging. Several computational methods have been developed to model the folding of RNA 3D structures and their interactions, mainly with proteins. However, these computational methods are nearing their limits, especially when the biological implications demand calculations of dynamics beyond a few hundred nanoseconds. For researchers facing such challenges, a more effective approach is to use coarse-grained modeling.

I will present strategies for computational modeling of RNA 3D structures and their interactions with other molecules. These strategies use a suite of methods from my laboratory, based on the SimRNA program. Our methods employ coarse-grained representations of molecules, utilize the Monte Carlo method for sampling conformational space, and use statistical potentials to approximate energy. They also help identify conformations that match biologically relevant structures. Specifically, I will discuss computational methods to determine RNA structure using low-resolution experimental data, such as chemical probing and electron microscopy.

References

Boniecki et al. Nucleic Acids Res. 2016 doi: 10.1093/nar/gkv1479
Ponce-Salvatierra, A. et al. Biosci. Rep. 2019 doi: 10.1042/BSR20180430
de Moura TR et al. Nucleic Acids Res. 2024 doi: 10.1093/nar/gkae144

SimRNA-sol: A novel method for simulating RNA folding with water and ions, and predicting tightly-bound solvent molecules in RNA 3D structures

Masoud Amiri Farsani[1] Filip Stefanik[1] Seyed Naeim Moafinejad[1] Janusz M. Bujnicki[1]

[1] *International Institute of Molecular and Cell Biology in Warsaw*

RNA molecules are essential to numerous biological processes, with their three-dimensional structures being fundamental to their functions. Metal ions are important for RNA folding, stability, and catalytic activity by stabilizing RNA architecture, mitigating negative charge repulsion, and facilitating catalytic reactions. Obtaining high-resolution experimental RNA structures is difficult, with 3D structures only available for a limited number of RNA molecules. Computational methods for RNA 3D structure prediction and modeling are therefore essential for linking sequence information to biological function. To tackle these challenges, we developed SimRNA-sol, an extension of SimRNA, which predicts RNA-solvent interactions with high accuracy using statistical potentials. We created a non-redundant training dataset of RNA structures containing tightly bound water molecules and the prevalent ions Mg^{2+}, K^{+}, and Na^{+}, all resolved by X-ray crystallography at resolutions better than 3.0 Å. SimRNA-sol simulations identified binding pockets for Mg^{2+}, K^{+}, Na^{+} ions and water molecules in a testing set of 65, 14, 10, and 35 structures, respectively. Our findings showed that including K^{+} ions significantly enhances the prediction of G-quadruplex structures. The accuracy of predicting binding sites for Mg^{2+}, Na^{+}, K^{+}, and water molecules was 65%, 53%, 75%, and 56%, respectively. Reference: Boniecki et al., Nucleic Acids Res. 44(7) (2016) e63. F. Leonarski et al Nucleic Acids Res, 45 (2017) 987-1004.

Session 2

Structural and functional diversity across large protein databases

Paweł Szczerbiak[1] Łukasz Szydłowski[1] Witold Wydmański[2] Tomasz Kosciolek[1]

[1]*Sano Centre for Computational Medicine, Kraków, Poland*

[2]*Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland*

Since the advent of AlphaFold and ESMFold, we have gained access to huge databases of protein structure predictions (over 800 million) that have surpassed the capabilities of traditional tools used till now to analyze a relatively small set of models (approx. 200,000) deposited in the Protein Data Bank. This catalyzed the development of highly scalable tools such as Foldseek, Foldcomp, and ProteStAr, enabling us to explore the full potential of the AlpahFold database (AF-DB) and ESMAtlas. Our work comprehensively examines the structural clusters obtained from the AlphaFold-Database (AF-DB), a high-quality subset of ESMAtlas (hclust30), and the Microbiome Immunity Project (MIP). While AF-DB, based on Uniprot, consists of protein structure predictions extensively studied over the years (with a significant eukaryotic component), the latter two databases contain only bacterial and archaeal proteins derived mainly from metagenomic studies. This provides a unique opportunity to understand the differences between these realms regarding their structural diversity. In addition, we elucidate the functional coverage of these databases by showing divergences and finding functional blind spots. Finally, we identify regions in the structural space that still need to be explored (annotated) because they are not well represented in Uniprot (AF-DB). Our findings lay the groundwork for more in-depth studies concerning protein sequence-structure-function relationships, where various biological questions can be asked concerning taxonomic assignments, environmental factors, or functional specificity, to name a few.

Strain-resolved metagenomics in analysis of gut microbiota dynamics within families

Katarzyna Sidorczuk[1,2] Ezgi Ozkurt[1,2] Klara Cerk[1,2] Anthony Duncan[1,2] Falk Hildebrand[1,2]

[1] *Quadram Institute Biosciences, Norwich Research Park, Norwich, United Kingdom*
[2] *Earlham Institute, Norwich Research Park, Norwich, United Kingdom*

The gut microbiota is a crucial factor in human health. Our gut microbiota is seeded at birth, starting with mostly facultative aerobes in the first weeks and gradually changing to anaerobes. Previous studies have shown that most of the early colonizing species are transmitted from mothers and that the diversity of gut microbiota increases as the children get older. Moreover, higher strain transmission rates have been associated with both kinship and cohabitation.

We investigated factors crucial for bacterial colonization and analysed strain transmission patterns between family members at unprecedented scale. We collected almost 5,000 publicly available shotgun metagenomic samples from 21 studies representing families and mother-child pairs. We processed the samples with MATAFILER pipeline developed in the group to assemble the metagenomes, predict genes and create a gene catalogue, bin them into metagenome assembled genomes (MAGs), and perform functional annotations.

We identified over 1,800 metagenomic species (MGS) with high-quality MAGs. More than 90% of these were resolved intra-specifically to track strain transmissions within family members and investigate their stability over time based on longitudinal studies. The species with most persistent strains belonged to Bacteroides, Alistepes and Parabacteroides genera. The highest transmission rates within families were observed for child-mother and child-child pairs. We further analysed de novo reconstructed pangenomes for each MGS to identify genes overrepresented in infants that could indicate their involvement in colonization process. We identified over a thousand of genes among 92 MGS with significant differences in distribution between age categories.

Our analysis of strain-resolved provides novel insights into strain dynamics within families and identifies genes characteristic of strains overrepresented in infants, potentially indicating factors that contribute to successful colonization.

MultiEM: Modelling Chromatin Structure from Nucleosomes to Chromosomal Territories

Sevastianos Korsak[1,2] Dariusz Plewczynski[1,2]

[1] *Faculty of Informatics and Mathematics, Technological University of Warsaw*
[2] *Center of New Technologies, University of Warsaw, Poland*

We present MultiEM, an advanced computational engine leveraging OpenMM for the detailed modeling of chromatin interactions, encompassing scales from nucleosomes to chromosomal territories. MultiEM requires minimal input data, including chromatin loops, compartments or subcompartments, and ATAC-Seq data. The engine operates by minimizing a Hilbert curve structure through a sophisticated multi-scale forcefield approach. This method incorporates harmonic bond interactions to model chromatin loops, block-copolymer interactions for compartmental dynamics, and lamina-associated interactions to simulate B compartment-lamina binding.

MultiEM is optimized for various computational environments, including laptops, workstations, and HPC clusters, ensuring rapid generation of chromatin structures within minutes to hours, depending on the desired granularity. The robustness of MultiEM is demonstrated across a range of input data scenarios, including Hi-C, ChIA-PET, Hi-ChIP, and single-cell data. Our results underscore the utility and efficiency of MultiEM in providing high-resolution insights into chromatin architecture, making it a valuable tool for genomic research and bioinformatics applications.

Computational modeling of chromatin three-dimensional structure from super-resolution microscopy

Zofia Parteka-Tojek[1,2] Jacqueline Jufen Zhu[3,4] Byoungkoo Lee[3] Karolina Jodkowska[2,5] Ping Wang[3] Jesse Aaron[6] Teng-Leong Chew[6] Krzysztof Banecki[1] Dariusz Plewczyński[1,2,3] Yijun Ruan[3,4]

[1] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*
[2] *Centre of New Technologies, University of Warsaw, S. Banacha 2c, 02-097, Warsaw, Poland*
[3] *The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT, 06030, USA*
[4] *Department of Genetics and Genome Sciences, University of Connecticut Health Center, 400 Farmington Avenue, Farmington, CT, 06030, USA*
[5] *Centre for Advanced Materials and Technologies, Warsaw University of Technology, Poleczki 19, 02-822, Warsaw, Poland*
[6] *Advanced Imaging Center, Janelia Research Campus, Howard Hughes Medical Institute, 19700 Helix Drive, Ashburn, VA, 20147, USA*

The three-dimensional genome structure is crucial for gene regulation and cellular functions. Advances in 3D genomics have uncovered fundamental chromatin folding structures known as chromatin loops—long-range spatial interactions mediated by protein factors. This study introduces ChromoLooping, a tool developed to model the 3D structure of chromatin using super-resolution microscopy data and to reconstruct chromatin loops from iPALM microscopy images. We focused on visualizing a highly frequent 13 000 base pairs (bp) long chromatin loop in GM12878 cells, identified through ChIA-PET and Hi-C techniques, using iPALM. We achieved a resolution of up to 3 nm x 3 nm x 2 nm in microscopical images and 120bp in the next generation sequencing data, providing detailed insights into the chromatin loop's folding in single cells. Our analysis collected thirteen high-quality images of the target chromatin region. We processed these images to precisely determine the positions of oligo probes attached to the chromatin fiber, achieving a localization precision of up to 2 nm and simulating a genomic resolution of 10bp. We also developed a method to generate simulated iPALM images from population 3D genomics data (Hi-C) to validate ChromoLooping's chromatin path reconstruction accuracy. This validation confirmed that ChromoLooping could accurately reconstruct loop structures given adequate oligo probe coverage. Comparing the physical distances in our image models with contact frequencies from ChIA-PET and Hi-C, we found a complex chromatin folding in the target region, suggesting that chromatin looping is influenced by multiple factors beyond protein-mediated anchors. Despite the variability observed between individual cells and populations, we found concordance between microscopic visualization and genomic data, as evidenced by the average distances map.

Furthermore, we demonstrated ChromoLooping's applicability in modeling chromatin structures from publicly available multi-domain light microscopy images and its adaptability for electron microscopy. We aimed to model chromatin conformation from 3D images represented as density fields, modifying the original method to create DNA paths between 3D within these fields. We also propose a method for identifying data points from 3D density fields. Our study provides novel insights into chromatin folding in vivo, despite limitations such as a small sample size, single-color DNA staining, and the absence of non-looping controls.

Microbiome Diversity and Intra-Community Synergies Across Climate-Defined Ecological Niches

Dagmara Błaszczyk[1,2] Krzysztof Mnich[3] Alina Frolova[1] Katarzyna Kopera[1] Kinga Zielińska[1] Witold Wydmański[1,2] Michał B.Kowalski[1,2] Renata Zbieć-Piekarska[4] Wojciech Branicki[1,5] Witold Rudnicki[3,6] Paweł P. Łabaj[1]

[1] *Jagiellonian University, Malopolska Centre of Biotechnology, Krakow, Gronostajowa 7a*
[2] *Jagiellonian University, Doctoral School of Exact and Natural Sciences, Krakow, Lojasiewicza 11*
[3] *University of Białystok, Computational Center, Bialystok, Konstantego Ciołkowskiego 1M*
[4] *Central Forensic Laboratory of the Police, Warszawa, Al. Ujazdowskie 7*
[5] *Jagiellonian University, Faculty of Biochemistry,Biophysics and Biotechnology, Krakow, Gronostajowa 7*
[6] *University of Białystok, Institute of Computer Science, Bialystok, Konstantego Ciołkowskiego 1M*

Microbiota research now focuses on exposome factors and metadata, identifying niche-specific microorganisms. Unlike most studies treating microorganisms separately, we explore their synergies and impact on classifying samples within Polish microclimate clusters. In our study, we use 949 soil samples collected from different locations in Poland in four seasons within a year. Samples have been sequenced with an extreme depth of over 120M paired-end reads. The sampling locations were selected based on climate characteristics supported by over 20 years of history of weather conditions parameters and represented three different Polish microclimate clusters. In our analysis, we used two tools: Kaiju and Kraken to obtain OTUs (Operational Taxonomic Units), which then were used as a feature table in MDFS (MDFS (MultiDimensional Feature Selection). MDFS is based on Mutual-information theory, to reveal synergies between microorganisms in corresponding climate niches. This further allows us to investigate how exploiting microbial synergies impacts the classification of samples into specific climate clusters. The comparison between Kaiju and Kraken showed that Kraken classified over 10,000 more OTUs than Kaiju, which results from the method those tools use. We confirmed the existence of microclimate-related and local-specific microbial communities, which aligns with earlier studies of MetaSUB Consortium on a global scale. Using features selected as synergistic we can obtain good results compared to other feature selection methods with visible reduction of features in less time and computational resources consuming process.

Sponsor

Generative AI in Health

Tomasz Żemojtel[1]

[1] *Amazon Web Services*

In a rapidly evolving landscape where the state of the art for generative AI changes weekly, the healthcare sector stands to benefit significantly from these advancements. Recent continuous developments, such as the release of Claude 3.5 by Anthropic, have demonstrated the potential of generative AI to outperform previous models like GPT-4 and Gemini 1.5, while dramatically reducing costs. However, to fully leverage the capabilities of GenAI, healthcare organizations need a robust data strategy encompassing data access, pipelines, and repository governance. Immediate applications, such as clinical documentation, speech recognition, and data structuring, have already shown remarkable results in improving efficiency and care quality. Furthermore, integrating GenAI with healthcare-specific data sets enables organizations to create tailored applications that deliver greater business value and patient outcomes. Examples include intelligent chatbots, automatic code generation, and summarizing complex medical narratives. In research, GenAI streamlines data management and accelerates clinical trial matching, while for patient engagement, it provides innovative tools for self-triage and personalized communication. The future of healthcare lies in adopting these transformative technologies, driven by data unique to each organization's needs and capabilities. This talk will explore the strategic approaches and real-world applications of GenAI in health, highlighting both current trends and future possibilities.

Posters

A Novelty Approach for De Novo Assembly of Current Data from Nanopore Sequencing -

Wiktor Kuśmirek[1]

[1] *Warsaw University of Technology, Institute of Computer Science*

Nanopore sequencing relies on detecting changes in the electrical current as a DNA strand translocates through a nanopore. These current variations are then translated into nucleotide sequences through a process known as basecalling, which employs various algorithms based on artificial intelligence. The output of basecalling is a set of nucleotide sequences, referred to as DNA reads. These reads are subsequently de novo assembled into longer DNA sequences. We propose a novel approach that bypasses the traditional reliance on DNA reads. Instead, our method directly compiles the raw current value changes into a long sequence, which is then converted into the final nucleotide sequence. In our work, we present the results of proof of concept research and preliminary conclusions.

ClaRNP: a classifier of contacts in 3D structures of RNA-protein complexes

Grigory Nikolaev[1] Eugene F. Baulin[1] Janusz M. Bujnicki[1]

[1] *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland*

RNA-protein interactions are fundamental to various cellular processes, including transcription, translation, splicing, and RNA modification. These interactions are characterized by their specificity and affinity, which are determined by hydrogen bonding, van der Waals forces, and electrostatic interactions. A comprehensive understanding of these interactions at the molecular level is essential for deciphering the mechanisms of RNA-protein recognition and function. We have developed a classifier for amino acid-nucleotide/nucleoside interactions ClaRNP, which takes as input 3D structures and annotates them with pairwise contacts that distinguish stacking interactions (involving the faces of nucleobases), pseudopairs (involving the Watson-Crick, Hoogsteen, or sugar edge of the base), and phosphate and ribose interactions for on the nucleotide/nucleoside side, and side-chain and backbone interactions on the amino acid side. According to the benchmarks with programs DSSR and fingeRNAt, ClaRNP can recognize and classify more nucleotide-amino acid doublets. This is because DSSR and fingeRNAt can classify only hydrogen bonds, van der Waals forces and stacking, and ClaRNP is based on geometric approach. ClaRNP compares the input doublet with the prepared database of doublets extracted from the experimentaly solved RNP structures. The classifier can be easily extended to include new types of spatial relationships between pairs or larger assemblies of nucleotide-amino acid residues.

SQUARNA - an RNA secondary structure prediction method based on a greedy stem formation model

Davyd R. Bohdan[1] Grigory I. Nikolaev[1] Janusz M. Bujnicki[1] Eugene F. Baulin[1]

[1] *International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4 02-109 Warsaw, Poland*

Non-coding RNAs play a great variety of roles in many cellular processes with their spatial structure known to dictate the functioning. RNA secondary structure largely determines the molecule's global fold, which together with the paucity of experimentally determined RNA 3D structures makes its knowledge crucial for determining the function of the molecule. Currently, there is no one good solution for de novo RNA secondary structure prediction, with the existing methods getting more and more complicated without substantial progress in accuracy and being subject to drastic limitations such as ignoring pseudoknots. In this work, we present SQUARNA, a new approach for de novo RNA secondary structure prediction based on a straightforward greedy stem formation model overcoming many limitations of the existing tools. The benchmarks show that SQUARNA is on par with state-of-the-art methods for a single sequence input and significantly outperforms existing tools for a sequence alignment input.

Reference-free ranking method for RNA 3D models

Tomasz Zok[1] Jan Pielesiak[1] Maciej Antczak[1,2] Marta Szachniuk[1,2]

[1]*Institute of Computing Science, Poznan University of Technology, Poznan, Poland*
[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland*

Interest in predicting RNA 3D structures has surged as more researchers recognize RNA's pivotal role in biology. Understanding RNA's complex structures and functions can revolutionize fields from genetics to drug design. Building on the successes of protein structure prediction, we now have the opportunity to apply similar advancements to RNA prediction.

One of the significant hurdles in this realm is evaluating the quality of the generated models. With modeling software often churning out numerous models per input—sometimes thousands—selecting the most accurate ones becomes essential. Traditionally, researchers have relied on energy calculations via force fields or statistical potentials; lower energy models are presumed more accurate. Yet, the energy landscape is fraught with local minima, making these results often ambiguous.

We propose an innovative solution: ranking multiple 3D RNA models by analyzing their base pairings and stacking interactions. By constructing a consensus secondary structure from this data, we can rank each model's interaction network against this benchmark, leading to a more definitive ranking system.

Our method was rigorously tested on publicly available RNA 3D modeling datasets, and the results demonstrated its superiority over traditional energy-based evaluations. This new approach promises to enhance the accuracy and reliability of RNA structure predictions, providing a valuable tool for researchers across the scientific spectrum.

Machine learning in biomodeling and evaluating the beneficial and adverse effects of using virtual environments in computational neuroscience

Beata Sokołowska[1] Ewa Sokołowska[2] Teresa Sadura-Sieklucka[3]

[1]*Bioinformatics Laboratory, Mossakowski Medical Research Institute, Polish Academy of Sciences, Warsaw, Poland*
[2]*Department of Developmental Psychology, Faculty of Social Sciences, Institute of Psychology, The John Paul II Catholic University of Lublin, Lublin, Poland*
[3]*Department of Geriatrics, National Institute of Geriatrics, Rheumatology and Rehabilitation, Warsaw, Poland*

Machine learning (ML) is an AI type focused on building computer systems that learn from data. ML is a powerful tool for solving problems, streamlining various complex operations, and automating tasks, which is used in many fields. ML offers a wide range of techniques, such as decision trees, rule induction, neural networks, support vector machines, clustering and classification methods, association rules, feature selection procedures, visualization, graphical models, or genetic algorithms. Many neuroscientists use innovative technologies (e.g., extended reality, XR) and the Internet of Things (IoT) with advanced computational algorithms in their research. Likewise, we use cutting-edge bioinformatics tools and novel virtual environments (VEs) in our modeling studies with healthy participants and in VR programs for various patient groups. The current research results show that VEs have high ecological value (in neuropsychology, neuropedagogy, neuropsychiatry, neurology or neurogeriatrics). Researchers, like our team, emphasize that ML algorithms have great potential, especially in computational neuroscience based on VR (XR). The report outlines that the use of VEs with ML (to evaluate the usability and effectiveness of various systems offered via VR) has especially potential in (a) more accurate VR (neuro)diagnostics, (b) more effective VR (neuro)therapy, and particularly (c) in promoting healthy aging. However, in addition to these beneficial effects, being in VEs can sometimes be associated with adverse symptoms of cybersickness. In interesting studies, ML models are used to identify unfavorable virtual symptoms early. In conclusion, ML (AI with IoT) and VEs in neuroscience and contemporary medical practice provide effective support for accurate diagnosis, evaluation of beneficial or adverse effects of therapy and rehabilitation, as well as natural aging processes, in the era of dynamic development of attractive and increasingly available IT/ICT technologies.

Magnetstein: a novel tool in qNMR and monitoring chemical reactions

Barbara Domżał[1] Magdalena Grochowska-Tatarczak[2] Krzysztof Kazimierczuk[2] Anna Gambin[1]

[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

[2] *Centre of New Technologies, University of Warsaw*

Monitoring chemical reactions using quantitative NMR spectroscopy can be a challenging task. Common difficulties include overlapping peaks of substrates and products, shifting of peak positions over time and distorted lineshapes due to magnetic field inhomogeneities. Here, we present Magnetstein: a novel tool for quantification of spectral components, offering solutions to these problems. The algorithm is based on the Wasserstein metric — a purely mathematical concept that has proven to be remarkably successful in the analysis of spectroscopic data. Magnetstein has already been widely tested on a task of estimating proportions of components in a mixture given the spectrum of this mixture and a library of components' spectra [1, 2]. Presently, we demonstrate its effectiveness in monitoring chemical reaction dynamics as well.

[1] Domżał, B., Nawrocka, E.K., Gołowicz, D., Ciach, M.A., Miasojedow, B., Kazimierczuk K. & Gambin, A. (2023). Magnetstein: An Open-Source Tool for Quantitative NMR Mixture Analysis Robust to Low Resolution, Distorted Lineshapes, and Peak Shifts. Analytical Chemistry. 96. 10.1021/acs.analchem.3c03594.

[2] https://github.com/BDomzal/magnetstein

Quantification of protein active sites evolvability with complexity flow graphs

Mateusz Twardawa[1,2] Piotr Formanowicz[1]

[1] *Institute of Computing Science, Poznan University of Technology*

[2] *ICT Security Department, Poznan Supercomputing and Networking Center*

Evolvability, the capacity of a biological system to generate heritable phenotypic variation, is a key factor in the adaptive potential of proteins. Understanding the evolvability of protein active sites is crucial for insights into enzyme function, adaptation and engineering. This study presents new method for quantification of structural and functional dynamics that contribute to the evolvability of protein active sites. The proposed method is based on complexity flow graph (CFG), that is a new structure extending concept of a fitness landscape. CFGs are able to track parameters related to protein site complexity relative to difficulty of catalyzing specific reaction starting from a defined protein sequence. In other words, the CFG represent information gain gradients, that can be observed during biological evolution or stochastic optimization. Since, protein active sites can be successfully modelled with cellular automata, this model was used to test application potencial of CFG in computational setting. The results show that CFGs are able to detect patterns that can be used in practice, enabling higher level of control during evolutionary experiments. In conclusion, CFG may possess significant implications for the fields of enzyme engineering and synthetic biology. Better understanding of evolvability of active sites, may lead to development of new frameworks and computational tools for designing robust and adaptable enzymes, thanks to evolutionary trajectory prediction and engineering.

Feature screening and model regularization for k-mer representations of biological sequences

Krystyna Grzesiak[1,2] Jakub Kołodziejczyk[2] Jarosław Chilimoniuk[2] Małgorzata Bogdan[1] Michał Burdukiewicz[2]

[1] *Faculty of Mathematics and Computer Science, University of Wrocław*

[2] *Clinical Research Centre, Medical University of Wrocław*

Proteins and peptides function effectively because they adopt specific spatial conformations. However, experimentally determining these structures is both costly and time-consuming. To accelerate this process, computational methods are often employed, leveraging the readily available amino acid sequences of proteins. Training deep learning models on such data, however, necessitates a large volume of annotated sequences. For peptide-specific models, where experimentally labeled sequences are scarce, classical statistical methods are preferred. Unfortunately, current state-of-the-art methods for peptide property prediction often fail to fully utilize the information embedded in amino acid sequences due to inefficient feature representations.

Classical methods like Generalized Linear Models (GLMs) require structured data, leading us to focus on k-mer representations of proteins. These representations can be formatted as binary sequences indicating the presence of consecutive k-mers, or as integer count sequences. The protein properties we aim to predict span various data types, including binary (e.g., presence of disordered regions), categorical (e.g., subcellular location), and continuous (e.g., minimum inhibitory concentration of antimicrobial peptides).

Our objective is to clarify the relationship between these protein properties and their k-mer representations. To achieve this, we introduce an advanced data simulation framework and methods for motif identification from k-mer data using information criteria like mBIC2 and regularization models. We conducted comprehensive benchmarking of various feature selection techniques, including Fast Correlation-Based Filter Solution (FCBF) and QuiPT, in conjunction with GLM-based regularization techniques. Our benchmarks identified the optimal feature selection methods for high-dimensional k-mer data, establishing a robust framework for accurately predicting protein properties and detecting motifs.

DNA tagging for physical object labeling

Anna Wąsowska[1] Tomasz Ociepa[1,2] Marek Miśkiewicz[3,1] Adam Kuzdraliński[1]

[1]*Department of Bioinformatics, Polish-Japanese Academy of Information Technology, Warsaw, Mazowieckie, 02-008, Poland, adamkuzdralinski@gmail.com*
[2]*Institute of Plant Genetics, Breeding and Biotechnology, University of Life Sciences in Lublin, Akademicka 15, 20-950 Lublin, Poland*
[3]*Institute of Computer Science, Maria Curie-Sklodowska University, Akademicka 9, 20-033, Lublin, Poland*

In recent years, numerous technological solutions for object labeling have been developed, characterized by varying levels of complexity and utility. Classic methods, such as barcodes and QR codes, offer straightforward data encoding but suffer from physical damage and limited range. More advanced technologies like radio frequency identification (RFID) and ultrawideband real-time location systems (UWB RTLS) provide enhanced capabilities but face challenges such as signal interference and energy consumption.

DNA-based techniques present a compelling alternative for object labeling. Originating in the 1980s, these methods leverage nucleic acids for tagging, as evidenced by patents like WO1987006383 and WO1990014441. Recent innovations, such as the 'DNA-of-Things' (DoT) framework by Koch et al.(1) and nanopore-based reading techniques by Doroschak et al.(2), have propelled the field forward. Companies like Haelixa, SelectaDNA, and Applied DNA Sciences have commercialized DNA tagging techniques, demonstrating their practical viability. DNA tagging is applicable to a wide range of materials, including fabrics, polymers, and metals, with stability dependent on surface characteristics and environmental factors.

The Department of Bioinformatics at the Polish-Japanese Academy of Information Technology, established on April 1, 2024, conducts research on DNA tagging (3) aimed at developing a system superior to those currently available or described in the literature. Future plans include the significant enhancement of DNA data encoding algorithms, the development of methods for embedding DNA in materials such as automotive paint, and the investigation of the physicochemical properties of DNA to enhance tag security against counterfeiting.

Koch, J. et al. Nat. Biotechnol. 38, 39–43 (2020). Doroschak, K. et al. Nat. Commun. 11, 5454 (2020). Kuzdraliński, A., et al. Nat. Commun. 14(1), 6052 (2023).

DEAD box helicase 5 regulates the three-dimensional organisation of heterochromatin upon exit from pluripotency

Misbah Abbas[1] Bondita Dehingia[1] Andrzej Szczepankiewicz[2] Aleksandra Piotrowska[1] Debadeep Chaudchury[1] Marcin Janowski[1] Hanna Nieznanska[2] Aleksandra Pękowska[1]

[1] *Dioscuri Centre for Chromatin Biology and Epigenomics, Nencki Institute of Experimental Biology, Polish Academy of Sciences, 3 Pasteur Street, 02-093 Warsaw, Poland*
[2] *Laboratory of Electron Microscopy, Nencki Institute of Experimental Biology, Polish Academy of Sciences, 3 Pasteur Street, 02-093 Warsaw, Poland*

DEAD-Box Helicase 5 (Ddx5) is an ATP-dependent RNA helicase that contributes to the regulation of transcription, splicing, and microRNA processing, at least in part due to its capacity to fold RNA. Ddx5 has been shown to restrain reprogramming suggesting its role as a gatekeeper of cell identity. Likewise, Ddx5 has been shown to interact with factors essential for proper development including CTCF that structures the three-dimensional (3D) folding of chromatin. However, the role of Ddx5 in cell differentiation is poorly understood. Here, we find that Ddx5 presence is essential for the proper development of the ES cells including neural stem (NS) cells. Using RNA-seq and DESeq2, we observed a minor effect of Ddx5 removal on the ES cell transcriptome, while NS cells derived from the Ddx5-/- ES cells feature broadly altered gene expression programs including numerous deregulated and misplaced loci. Hence, we provide evidence of a unique role of Ddx5 in shaping the transcriptome in the lineage-committed cells. Moreover, using rMATS, we identified exons featuring an altered pattern of splicing upon the removal of Ddx5. There were more retained than skipped exons in the Ddx5-/- NS cells. Exons aberrantly included in Ddx5-/- NS cells had weaker splice sites, and were flanked by larger introns than exons that were more frequently excluded in the absence of Ddx5. In addition, our data shows loci affected by Ddx5 loss frequently relate to heterochromatin assembly. Electron microscopy analysis confirmed the loosening of heterochromatin in the Ddx5-/- NS cells. Loss of heterochromatin clustering was accompanied by an increased transcriptional activity of simple repeats. Collectively, this data reveals a previously unappreciated heterochromatin structuring role of Ddx5 in development. Altogether, Ddx5 exerts a differentiation-stage-specific role on the transcriptome.

Understanding the LC3/PEBP1 complex in cell death mechanisms: in search of binding spots, interaction dynamics and spatial structure

Julia Duda[1] Karolina Mikulska-Rumińska[1]

[1] *Nicolaus Copernicus University in Torun, Poland*

PE-binding protein 1 (PEBP1) is a multitask and versatile protein present in the human body that takes part in or even initiates various cell death programs like ferroptosis, necroptosis or autophagy [1]. An incorrect course of the degradation pathway mechanisms can lead to various illnesses, which is why the comprehension of the PEBP1 protein is crucial regarding its possible binding partners. Since the details of the various degradation pathway mechanisms are still unclear, further studies are conducted to understand the cell death mechanisms in hopes of understanding, and potentially inhibiting or activating them as needed (e.g. in therapies or drug design). The study was conducted to obtain the spatial structure of the LC3/PEBP1 complex, which initiates the autophagy pathway that plays an essential role in the maintenance of cell homeostasis in the human body. Understanding the mechanism of this interaction is essential for further research regarding possible activation or inhibition of said pathway and therefore expanding our knowledge of the versatility of the PEBP1 protein. Molecular dynamics simulations, docking, and the PRS method were used to understand the interactions between both proteins and their potential binding sites. The studies consisted of analyzing the LC3/PEBP1 complex - its behavior, most commonly interactive amino acid residues, and the stability of the complex. The obtained results indicate the successful determination of the spatial structure of the complex.

Acknowledgments: Work supported by Polish National Science Centre no. 2022/46/E/ST4/00053.

1. Andrew M. Lamade, et al., 2022, Inactivation of RIP3 kinase sensitizes to 15LOX/PEBP1-mediated ferroptotic death. Redox Biology, 50: 102232

QUBO formulation of Constrained Multiple Sequence Alignment

Katarzyna Nałęcz-Charkiewicz[1]

[1] *Warsaw University of Technology*

A novel formulation for constrained multiple sequence alignment (CMSA) as a quadratic unconstrained binary optimization (QUBO) problem is presented, utilizing the capabilities of quantum annealers (QA). Initially, an overview of existing QUBO formulations for multiple sequence alignment (MSA) is provided, followed by a discussion on incorporating constraints into these MSA formulations to adapt them for the CMSA task using two distinct methodologies. The computational complexity of these models is analysed, and their practicality for quantum annealing is assessed.

imputomics 2.0: comprehensive R package for missing data imputation in proteomics data

Jarosław Chilimoniuk[1] Krystyna Grzesiak[2,1] Jakub Kołodziejczyk[1] Adam Krętowski[1] Michał Ciborowski[1] Michał Burdukiewicz[1,3]

[1] *Clinical Research Centre, Medical University of Białystok, Białystok, Poland*
[2] *Faculty of Mathematics and Computer Science, University of Wrocław, Wrocław, Poland*
[3] *Institute of Biotechnology and Biomedicine, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain*

Background: Missing data is a common issue in proteomics, which can negatively affect downstream statistical analyses and the interpretation of results. Missing values (MV) in proteomics can be divided into two types, missing not at random (MNAR) and missing at random (MAR). It is widely believed that the third MV type, missing completely at random (MCAR), is the same as MAR. Various methods have been proposed to tackle this issue, from simple approaches like zero, mean, and median imputation to more complex techniques like random forest, singular value decomposition, and k-nearest neighbors. However, selecting the best method depends on the data type and research question.

Aim: We aim to provide a comprehensive overview of the existing algorithms for imputing missing values in proteomics datasets, considering different patterns of missingness, such as MAR, MNAR, and mixtures.

Methods: To address this issue, we systematically reviewed the literature on imputation methods for proteomics data. The selected imputation methods included basic techniques as well as advanced machine learning-based approaches.

Results: We intend to expand imputomics, our R package and Shiny web server originally developed for MV imputation in metabolomics datasets. We plan to incorporate identified and tested implementations of the MV imputation algorithms (MVIAs) that were used with proteomic datasets.

Datasets for benchmarking RNA design algorithms

Jan Badura[1] Tomasz Zok[1] Agnieszka Rybarczyk[1]

[1] *Poznań University of Technology*

RNA molecules play vital roles in many biological processes, such as gene regulation or protein synthesis. The adoption of a specific secondary and tertiary structure by RNA is essential to perform these diverse functions, making RNA a popular tool in bioengineering therapeutics. The field of RNA design responds to the need to develop novel RNA molecules that possess specific functional attributes. In recent years, computational tools for predicting RNA sequences with desired folding characteristics have improved and expanded. However, there is still a lack of well-defined and standardized data sets to assess these programs. Here, we present a large dataset of internal and multi-branched loops extracted from PDB-deposited RNA structures that encompass a wide spectrum of design difficulty. Furthermore, we conducted benchmarking tests of widely utilized open-source RNA design algorithms employing this data set.

The link between amyloid-related diseases and functional amyloids

Jakub Wojciechowski[1] Alicja Wojciechowska[2] Kinga Zielińska[3] Johannes Soeding[4] Tomasz Kościółek[1] Małgorzata Kotulska[2]

[1] *Sano Centre for Computational Medicine, Cracow, Poland*
[2] *Wroclaw University of Science and Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Wroclaw, Poland*
[3] *Malopolska Centre of Biotechnology UJ, Cracow, Poland*
[4] *Max Planck Institute for Multidisciplinary Sciences, Gottingen, Germany*

Amyloids are insoluble protein aggregates with a cross-beta structure which are traditionally associated with neurodegeneration. However, similar structures, named functional amyloids, play important roles in biofilm stabilization and construction, cell signaling, and even melanin production. Multiple examples of amyloid-producing bacteria inhabit the human microbiome. Given the by now undisputed link between health and gut microbiota, in this work, we computationally analyze the human microbiome in the context of functional amyloids. We show their grand diversity in gut microbiota across different taxonomic units and frequent presence in the extracellular space. We predict interactions between gut microbiome functional amyloids and human proteins and observe their potential to trigger inflammation and affect transport and signaling processes, similarly to pathological amyloids. Finally, we perform a metagenomic study and find a greater abundance of functional amyloids in diseased patients than in healthy controls. Our results provide a rationale for the suspected existing link between amyloid-related diseases and functional amyloids.

Identifying Essential Therapeutic Targets for Hyperglycemia-Induced Atherosclerosis Treatment Using a Petri Net-Based Model

Agnieszka Rybarczyk[1,2] Dorota Formanowicz[3] Piotr Formanowicz[1]

[1] *Institute of Computing Science, Poznan University of Technology*

[2] *Institute of Bioorganic Chemistry, Polish Academy of Sciences*

[3] *Poznan University of Medical Sciences, Department of Medical Chemistry and Laboratory Medicine*

Persistent high glucose levels, a hallmark of diabetes mellitus (DM), lead to widespread cellular damage. Atherosclerosis, which arises in conjunction with glucose metabolism disturbances, significantly complicates the condition and progresses more rapidly in DM, making it crucial to slow its advancement. Identifying therapeutic targets within the intricate network of molecules and processes, without disrupting essential functions, is challenging. Computational methods, such as in silico analysis, are vital for effectively identifying potential therapeutic targets. In our study, we utilized a Petri net model to pinpoint key network nodes whose inhibition could hinder atherosclerosis in the context of hyperglycemia. Our results indicate that inhibiting isoforms of protein kinase C in diabetic patients could help slow the progression of atherosclerosis. Additionally, we found that inhibiting aldose reductase could decelerate atherosclerosis and reduce PKC expression in DM. Targeting oxidative stress by inhibiting the AGE-RAGE axis emerges as a promising therapeutic strategy for managing hyperglycemia-induced atherosclerosis. Although blocking NADPH oxidase, the primary enzyme responsible for generating reactive oxygen species (ROS) in blood vessels, only slightly impeded atherosclerosis development, it effectively halted the increased production of mitochondrial ROS associated with mitochondrial dysfunction. These findings provide a foundation for more in-depth studies.

SimDNA: a coarse-grained method for DNA folding simulations and 3D structure prediction, and for the sampling of conformational landscapes involving canonical and non-canonical structures

Maciej Maciejczyk[1,2] S. Naeim Moafinejad[1] Michal J. Boniecki[1] Janusz M. Bujnicki[1]

[1] *Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland*
[2] *Department of Physics and Biophysics, University of Warmia and Mazury in Olsztyn, ul. Oczapowskiego 4, 10-719 Olsztyn, Poland*

DNA, the blueprint of life, primarily forms a double helix but can also create structures like junctions, triplexes, and quadruplexes. These structures are essential for cellular functions, including gene expression regulation, replication, and genome stability maintenance.

Exploring DNA structure through methods like X-ray crystallography, NMR, and Cryo-EM spectroscopy is crucial but accompanied by challenges. These methods can be costly and time intensive. X-ray crystallography captures static snapshots of DNA conformations, lacking dynamic insights. Moreover, NMR is restricted in its ability to analyze smaller DNA molecules, while achieving high-resolution Cryo-EM density maps is more common for larger biomolecules, such as those with 150kDa.

SimDNA, a new computational tool, addresses these challenges. It predicts DNA 3D structures using a coarse-grained representation and Monte Carlo dynamics, a statistical method that efficiently explores possible conformations by sampling from the most probable states. This approach allows SimDNA to accurately fold various DNA forms, including duplexes, junctions, and non-canonical structures like triplexes and G-quadruplexes, even without external restraints.

Furthermore, SimDNA enables guided simulations using data from experiments or other computational methods, providing a versatile tool for researchers. This flexibility allows user-defined restraints to focus simulations on specific interactions or structural configurations, facilitating the study of transitions between different DNA structures. Overall, SimDNA holds great promise for advancing our understanding of DNA behavior, offering insights into fundamental biological processes and aiding in biomedical research and therapeutic development.

I will discuss the SimDNA representation, conformational sampling algorithm, scoring function, and present the preliminary results.

Proteomic profiling of S. cerevisiae strains perturbed in RNA polymerase III activity by SWATH mass spectrometry

Ignacy Makowski[2] Roza Szatkowska[1] Paweł Ciborowski[3] Anna Gambin[2] Malgorzata Adamczyk[1]

[1] *Laboratory of Systems and Synthetic Biology, Chair of Drugs and Cosmetics Biotechnology, Faculty of Chemistry, Warsaw University of Technology, Poland*
[2] *Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Poland*
[3] *Department of Pharmacology and Experimental Neuroscience, University of Nebraska Medical Center, Omaha, NE, USA*

The regulation of metabolism in *Saccharomyces cerevisiae* has been extensively studied, revealing significant insights into yeast's proteome homeostasis. This study aims to expand upon our previous findings that glycolytic flux in *S. cerevisiae* is dependent on RNA polymerase III and its negative regulator Maf1 by comparing the protein levels of the wild-type (WT) and mutant strains perturbed in RNA polymerase III (RNAP III) activity (*maf1*Δ and *rpc128-1007*) under fermentative and non-fermentative growth conditions.

Upregulation of non-coding RNA synthesis is associated with a decrease in TCA cycle activity, causing mitochondrial dysfunction. *rpc128-1007*, the strain that is unable to increase tRNA synthesis due to a mutation in the RNAP III subunit C128, shows increased TCA cycle activity under non-fermentable conditions. The primary objective of this study was to comprehensively understand how the absence of MAF1 and the point mutation in the RNAP III subunit influence mitochondrial fitness at the proteome level.

We collected proteomics data using data-independent, acquisition-based SWATH-MS. For SWATH-MS measurements, non-labeled protein samples were digested, and the resulting peptides were analyzed by liquid chromatography coupled to a tandem mass spectrometer operating in the data-independent acquisition (DIA) mode.

For data analysis, we utilized Panther for gene ontology (GO) analysis and DAVID for functional annotation clustering. Furthermore, Cytoscape was employed to construct and visualize the interaction networks, enabling us to discern significant pathway alterations and their potential biological implications. This research provides valuable insights into the metabolic adjustments and regulatory mechanisms in *S. cerevisiae*, the workhorse of the biotechnology sector.

Time Series Forecasting of Bacterial Taxonomy and Function via Deep Representations

Zuzanna Karwowska[1,2,3,4] Marcin Mozejko[5] Ewa Szczurek[5,6] Jaroslaw Biliński[4] Tomasz Kosciolek[3]

[1] *Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland*

[2] *Doctoral School of Natural and Exact Sciences*

[3] *SANO Centrum Zindywidualizowanej Medycyny Obliczeniowej*

[4] *Human Biome Institute*

[5] *MiM UW*

[6] *Institute AI for Health, Helmholtz Zentrum München*

The gut microbiome plays a crucial role in human health and disease. Longitudinal studies are increasingly recognized as essential to understanding its dynamic nature, offering insights that can lead to personalized health interventions like tailored diets and probiotics. However, the lack of dense time series data from healthy individuals complicates developing predictive models. Here, we propose a deep neural network that utilizes large cross-sectional metagenomics datasets to learn patterns (states) of the human gut microbiome taxonomy and function in the human population. These states are defined by deep embeddings—abstract representations of the subject's microbiome composition and functional potential and are used by a second model in order to predict how the gut microbiome changes in time. We include microbiome taxonomy with its functional potential due to higher functional overlap between individuals, as the same taxa can perform identical functions in different subjects. Our project leverages recent findings that despite its complexity, the gut microbiome exhibits predictable patterns. By using deep embeddings to simplify these patterns, we aim to improve the predictability of the microbiome's temporal changes. We show that our model is able to correctly embed a subject based on their gut microbiome and we can define drivers of these embeddings. We also show the use of our model on time series data to predict how the gut microbiome will change after fecal microbiota transplant. This framework suggests a promising direction towards a microbiome-centric foundation model for developing new therapeutic strategies based on microbiome dynamics.

Enhancing Amino Acid Classification in NMR Data Using Sliced-Wasserstein Optimal Transport

Zofia Gruba[1] Barbara Domżał[1] Błażej Miasojedow[1] Krzysztof Kazimierczuk[2] Anna Zawadzka-Kazimierczuk[3] Anna Gambin[1]

[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*
[2] *Centre of New Technologies, University of Warsaw, Warsaw, Poland*
[3] *Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Warsaw, Poland*

We implemented an algorithm that leverages the approximation of the multidimensional measure of optimal transport, specifically utilizing the sliced-Wasserstein distance, to classify amino acids from data obtained through NMR experiments. This analysis focused on a set of intrinsically disordered proteins identified within the BMRB database[1]. In previous approaches, linear discriminant analysis and linear regression methods were employed for this classification task. Our new method represents a significant advancement by incorporating the sophisticated and precise framework of optimal transport theory, which enhances the accuracy and robustness of amino acid classification in these complex protein structures. Our approach involves computing the transport between the training and test sets by randomly sampling a specified number of lines in the given multidimensional space, where the dimension is determined by the number of nuclei types involved in an NMR experiment. We project the data onto these sampled lines and, for each projection, perform the transport between the training and test sets in one dimension. Instead of using the Euclidean metric, classification is based on the k-nearest neighbors method, utilizing values derived from the transport plan. The main advantage of our approach is a different perspective on the classification problem and the use of the optimal transport metric, which has proven effective for other NMR data problems. Preliminary results are highly promising, indicating that our method achieves greater accuracy in classifying challenging amino acids, such as cysteine and glutamine, compared to linear discriminant analysis. 1 ""Biological Magnetic Resonance Data Bank"", J C Hoch; K Baskaran; H Burr; J Chin; H R Eghbalnia; T Fujiwara; M R Gryk; T Iwata; C Kojima; G Kurisu; D Maziuk; Y Miyanoiri; J R Wedell; C Wilburn; H Yao; M Yokochi; Nucleic Acids Research, Volume 51, Issue D1, 6/1/23, Pages D368–D376 doi: 10.1093/nar/gkac1050

Confident datasets of client, driver and negative proteins in liquid-liquid phase separation

Michał Burdukiewicz[1] Carlos Pintado-Grima[2] Oriol Bárcenas[2,3] Valentín Iglesias[1] Eva Arribas-Ruiz[2] Salvador Ventura[2]

[1] *Clinical Research Centre, Medical University of Białystok, Białystok, Poland*
[2] *Institut de Biotecnologia i de Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, Barcelona, Spain*
[3] *Institute of Advanced Chemistry of Catalonia (IQAC), CSIC, Barcelona, Spain*

Proteins self-organize in dynamic cellular contexts by assembling into reversible biomolecular condensates by liquid-liquid phase separation (LLPS). Such condensates can be constituted by one or many proteins, which may have different roles in the final supramolecular structure. While driver proteins form condensates by themselves without the requirement of any partner, client proteins are later recruited into them and are not essential for their integrity. Although several databases collect proteins undergoing LLPS, they differ in conceptual considerations. Consequently, a lack of data interoperability between resources hampers the integrative usage of their contents. Moreover, there is an apparent missing consensus on selecting proteins without any explicit experimental association with condensates (negative data). These two aspects have prevented the generation of specific confident predictors and fair benchmarks.

To alleviate this issue, in this work, we have explored protein information from all relevant LLPS databases to generate confident datasets of client and driver proteins through an integrated biocuration protocol. Besides, we introduce standardized negative datasets, including both globular and disordered proteins. To validate our datasets, we painstakingly investigated specific physicochemical properties related to LLPS in different sets of sequences. Interestingly, in the first analysis, we have observed that certain features can be used to distinguish drivers and clients of LLPS but also against negative proteins (non-participants of condensates). The datasets generated in this study are readily available at https://llpsdatasets.ppmclab.com and https://github.com/PPMC-lab/llps-datasets.

Our highly confident datasets are expected to be used to train a new generation of specific multilabel models for attaining better predictive performances, building more standardized benchmarks and avoiding sequential biases commonly observed for IDRs.

Statistical evaluation of existing CRISPR-Cas9 deep learning models predicting off-target activities

Maciej Powierża[1]

[1] *Hirszfeld Institute of Immunology and Experimental Therapy*

Since its discovery in 2012, CRISPR-Cas9 technology has been used with promising outcomes in numerous biological research studies ranging from plant genetic engineering and biofuel production to testing potential therapies for human genetic disorders and cancer. Unfortunately, Crispr-Cas9 is still of limited application in clinical context due to occurrence of off-target effects [1]. To mitigate this problem, various experimental methods like GUIDE-seq, SITE-seq, and Digenome-seq have been developed for off-target detection and validation. To reduce the associated costs, *in silico* algorithms have been introduced to predict potential off-target sites, narrowing the search scope and focusing experimental efforts on preselected sequences, thus further reducing expenses [4].

During the course of recent years three main classes of off-target predicting algorithms have been explored, including scoring-based, machine learning and deep learning models. Among these classes the deep learning models are at present of greatest interest due to their fast evolvement and high accuracy when compared with the other two classes. Up to now a considerable amount of papers accompanied with source codes have been published, presenting original models intended to predict off-target sites based on genomic input data. As the number of models increases, their evaluation becomes more and more important. So far, the metrics such as AUROC, AUPRC and Spearman correlation were the most frequently used by the authors dealing with the subject for model's evaluation and benchmarking [4].

We hypothesise, though, based on indications provided by some deep learning theorists [2][3], that the currently used evaluation measures may prove not optimal for reliable comparisons of existing models. Therefore, we propose a new evaluation based on additional performance measures such as Matthew's Correlation Coefficient (MCC) and Modified Confusion Entropy (MCEN) as well as on statistical tests performed on replicated evaluations of a model's predictive power.

References: [1] DOI: 10.1038/s41587-020-0561-9 [2] Demšar, 'Statistical Comparisons of Classifiers over Multiple Data Sets', J. Mach. Learn. Res., Dec. 2006. [3] DOI: 10.1162/089976698300017197 [4] DOI: 10.1093/bib/bbad131

Aggregating gut: on the link between amyloid-related diseases and functional amyloids - a computational study

Alicja Wojciechowska[1] Jakub Wojciechowski[2] Kinga Zielińska[3] Johannes Soeding[4] Tomasz Kościółek[2] Małgorzata Kotulska[1]

[1] *Wroclaw University of Science and Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Wroclaw, Poland*
[2] *Sano Centre for Computational Medicine, Cracow, Poland*
[3] *Malopolska Centre of Biotechnology UJ, Cracow, Poland [correct?]*
[4] *Max Planck Institute for Multidisciplinary Sciences, Gottingen, Germany [correct?]*

Amyloids are insoluble protein aggregates with a cross-beta structure which are traditionally associated with neurodegeneration. However, similar structures, named functional amyloids, play important roles in biofilm stabilization and construction, cell signaling, and even melanin production. Multiple examples of amyloid-producing bacteria inhabit the human microbiome. Given the by now undisputed link between health and gut microbiota, in this work, we computationally analyze the human microbiome in the context of functional amyloids. We show their grand diversity in gut microbiota across different taxonomic units and frequent presence in the extracellular space. We predict interactions between gut microbiome functional amyloids and human proteins and observe their potential to trigger inflammation and affect transport and signaling processes, similarly to pathological amyloids. Finally, we perform a metagenomic study and find a greater abundance of functional amyloids in diseased patients than in healthy controls. Our results provide a rationale for the suspected existing link between amyloid-related diseases and functional amyloids.

Investigating Antimicrobial Resistance in Poland: Seasonal and Geographic Analysis of Soil Resistome Markers Along the Vistula River

Rodolfo Brizola Toscan[1] Dagmara Błaszczyk[1] Katarzyna Kopera[1] Alina Frolova[2] Balakrishnan Subramanian[3] Paweł Łabaj[1]

[1] *Małopolska Centre of Biotechnology, Jagiellonian University*
[2] *Institute of Molecular Biology and Genetics of NASU, Kyiv Academic University, Kyiv, Ukraine*
[3] *Institute of Computer Science, University of Białystok*

Antimicrobial resistance (AMR) is a significant public health challenge driven by anthropogenic factors. This study investigates seasonal and geographic variations in soil resistome markers across Polish cities, aiming to identify region-specific resistance patterns for targeted AMR control. A total of 960 soil samples were collected from 80 locations, sequenced using Whole Genome Sequencing (WGS) on the Illumina NovaSeq 6000 platform. The analysis utilized AMR++ v3.0.0 for resistome marker detection and data normalization. The Pythagorean theorem was applied to compute an average resistome index, revealing significant correlations between resistome profiles, soil and water pH, and phosphate levels. The resistome index was observed to increase gradually for samples along the Vistula River, indicating contamination caused by anthropogenic influence. This comprehensive resistome profiling provides a baseline for future studies and interventions to mitigate AMR spread.

Investigation of the temporal changes in brain activity based on functional Magnetic Resonance Imaging data.

Patryk Mierzejewski[1] Krzysztof Kotlarz[1] Patryk Baran[1] Jakub Borysewicz[1] Patryk Brzezicha[1] Daniel Duszczak[1] Joanna Szyda[1] Magdalena Frąszczak[1]

[1] *Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw 51-631, Poland*

Functional Magnetic Resonance Imaging (fMRI) is an advanced technique for visualising and analysing brain activity in both spatial and time dimensions. In this study, fMRI data were collected from 88 patients (41 women and 46 men), each participating in three different conceptual scenarios: the Faces scenario, the Stroop Test scenario, and the Reappraisal scenario. The aim was to investigate the temporal changes in brain activity, which provides insights into how different cognitive tasks generate different neural responses. The Faces scenario comprised 176 images of faces expressing various emotions, the Stroop Test scenario comprised 396 images of colour words printed in opposite hues, and the Reappraisal scenario featured 570 images. Each 3D MRI image comprised 35 2D slices, with each slice made up of voxels, the smallest units of three-dimensional space. For each patient, MRI scans were pre-processed under each scenario, yielding a series of 3D images over time. To allow for further analysis, these temporally ordered 3D pictures were converted into 4D fMRI datasets using SPM12 software. Then statistical analysis containing Multidimensional Scaling, Multiple Correlation, Kolmogorov test, tests for variances and Wilcoxon sum of rank test were performed using R software. The study results showed significant differences in brain activity depending on the type of task and stimuli. The Stroop test and the Faces test engaged different areas of the brain, reflecting the specifics of cognitive and emotional information processing. The Stroop test was associated with greater activity in areas related to attention control and executive processes, such as the anterior cingulate cortex and prefrontal cortex. In contrast, the Faces test activated areas related to emotion processing, such as the amygdala and temporal cortex. Comparing the results of the first and second examination within each task revealed differences in brain activity patterns, suggesting adaptive brain mechanisms in response to repeated stimuli. These patterns may indicate learning and adaptation processes, where the brain modifies its activity in response to familiar tasks. For instance, repeated performance of the Stroop test could lead to more efficient processing of cognitive conflict, reflected in decreased activity in some brain areas during the second session. These results can contribute to a better understanding

of brain functioning in the context of different cognitive tasks.

Comparison of automatic cell labelling methods for single-cell RNA-seq data

Patrycja Rosa[1,2] Aleksander Jankowski[1]

[1] – *Laboratory of Molecular Neurobiology, Nencki Institute of Experimental Biology, Warsaw, Poland*
[2] – *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland*

The field of single-cell transcriptomics has tremendous potential for uncovering biological processes at the cellular level. The experimental design allows for sequencing of many cells at once while preventing batch effects caused by technical issues. It is even possible to sequence multiple experimental conditions in a single sequencing run, which streamlines the analysis process. In the computational analysis of the sequencing data, numerous tools have been developed to facilitate data cleaning and preprocessing before further analysis.

One major challenge in downstream analysis is properly labelling cell types based on prior biological knowledge. Over the years, many different algorithms have been developed to speed up this process, but they are far from perfection. In practice, these algorithms often struggle to accurately assign cell types when analysed cell populations are not very similar or are in a different stage of cell differentiation.

To investigate the question further, we benchmarked six available methods of automatic cell labelling . We evaluated the performance of these methods using publicly available human-annotated single-cell RNA sequencing datasets. We tested each method's performance within datasets in terms of accuracy, percentage of unclassified cells, and computation time. Also, we assessed the methods' sensitivity to input features, number of cells per population, and their performance across different annotation levels and datasets. Our findings indicate that most classifiers perform reasonably well on a variety of datasets, but their accuracy decreases for complex datasets with overlapping classes of annotations.

Healthy microbiome - moving towards functional interpretation

Kinga Zielińska[1] Klas I. Udekwu[2,3] Witold Rudnicki[4,5] Alina Frolova[6,7] Paweł P Łabaj[1]

[1]*Małopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland*
[2]*Department of Biological Sciences, University of Idaho, Moscow, ID 83843, U.S.A.*
[3]*Swedish Environmental Epidemiology Centre, Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala, SE75007, Sweden*
[4]*Faculty of Computer Science, University of Białystok, Białystok, Poland*
[5]*Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland*
[6]*Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, Kyiv, Ukraine*
[7]*Kyiv Academic University, Kyiv, Ukraine*

Microbiome-based disease prediction has significant potential as an early, non-invasive marker of multiple health conditions attributable to dysbiosis of the human gut microbiota, thanks in part to decreasing sequencing and analysis costs. Existing tools, or microbiome health indexes, are often based solely on microbiome richness and are heavily dependent on taxonomic classification. More recently, an ecological approach has led to increased understanding of microbiome, which reveals substantial restrictions of such approaches. In our study, we introduce a new health index created as an answer to updated microbiome definitions. The novelty of our approach is a shift from a traditional approach of phylogenetic classification, towards a more holistic consideration of metabolic function including ecological interactions between species in the effort to distinguish between healthy and diseased states. We compare this to not only the taxonomy-based Gut Microbiome Health Index (**GMHI**) and the high dimensional principal component analysis (**hiPCA**)method, the most comprehensive indices to date, but also to taxon- and function-based Shannon entropy and demonstrate a significant improvement to these approaches. We validate our index's performance using a variety of complementary benchmarking approaches on datasets representing a range of gut health conditions and showcase the robustness of its superiority over the **GMHI** and the **hiPCA**. Overall, we emphasize the potential of this approach and advocate a shift towards functional approaches in order to better understand and assess microbiome health as well as to provide directions for future index enhancements.

Factor analysis and correlated topic model for multi-modal data integration

Małgorzata Łazęcka[1,2] Kazimierz Oksza-Orzechowski[1] Ewa Szczurek[1,3]

[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*
[2] *Institute of Computer Science Polish Academy of Sciences*
[3] *Institute of AI for Health, Helmholtz Munich*

Integrating information across various data modalities can be beneficial for gaining valuable insights into underlying phenomena. Numerous methods exist for multi-modal data integration, ranging from linear matrix factorization-based approaches to nonlinear methods employing i.e. deep generative models. However, integrating data becomes particularly challenging when one or more modalities exhibit complex structures, such as image-based, or spatial. In such cases, existing strategies often rely on preprocessing structured data as an initial step. We present FACTM, a novel method that leverages a Bayesian probabilistic graphical model to address this challenge. Our approach combines two features. Firstly, it employs a multi-modal factor analysis (FA) to integrate information and identify common latent factors shared across all modalities, including structured data. Secondly, it uses a correlated topic model (CTM) to uncover the structure of the complex data. Specifically, the CTM part identifies meaningful clusters and shares information about the observation-wise changes of population fractions of specific clusters with the FA component of the model. Importantly, our model extracts information from complex modalities and runs factor analysis simultaneously, allowing both components of the model to potentially enhance each other's performance. In the context of spatial imaging of single cells, the CTM component of FACTM clusters spatial niches (groups of cells) into niche types. These niche types are defined as the distribution of cell types within them, and they are also inferred by the model. Meanwhile, the FA component integrates information from various spatial and non-spatial modalities to uncover common latent factors. We use the variational Bayesian framework with a mean field approximation. Optimal parameters are determined by maximizing the evidence lower bound (ELBO). In the poster, we will provide a description of the model, along with results demonstrating its practical application. Specifically, we will present findings derived from multi-omics data obtained from the IMMUcan consortium. The dataset contains spatial single-cell imaging modalities (IMC and mIF), along with non-spatial ones.

Distinctive Gut Microbiota Characteristics in Non-Food Allergy Patients: Insights from Multidimensional Feature Selection and Clique Analysis

Sajad Shahbazi[1] Piotr Stomma[1] Krzysztof Mnich[1] Witold Rudnicki[1]

[1] *Computational Center, University of Białystok*

Background: The role of intestinal microbiota in allergic conditions, including non-food allergies (NFA), has garnered attention, yet its specific influence on NFA in adults remains underexplored. This study aims to identify distinct gut microbiota characteristics associated with NFA and healthy individuals, utilizing advanced analytical methods such as multidimensional feature selection (MDFS) and clique analysis. Methods: Using data from the EMBL-EBI ENA project (PRJEB11419), we analyzed 6302 samples categorized into healthy individuals and NFA patients, comprising 1012 bacterial taxa after preprocessing. MDFS and random forest models identified bacterial communities distinguishing NFA from healthy groups. Taxa frequencies in cliques were analyzed to explore specific bacterial correlations. Statistical comparisons via t-tests and random forest models assessed bacterial community differences. Results: Non-food allergy (NFA) patients exhibited significantly higher bacterial diversity than healthy controls. Key genera contributing to these differences included Veillonella, Paucibacter, Kocuria, Burkholderiaceae, Leuconostoc, Neisseria, Streptococcaceae, Rothia, Actinomyces, Pasteurellaceae, Haemophilus, Acinetobacter, Betaproteobacteria, Streptococcus, Lactobacillus, and Fusobacterium. Network analysis revealed greater complexity in the NFA group, predominantly involving taxa from Firmicutes. In NFA conditions, four out of 319 cliques demonstrated strong associations. Random forest models showed high AUC values for total taxa frequency and abundances in these cliques across all taxonomic levels, each exceeding 0.9.

Conclusions: Gut microbiota composition and interactions vary significantly between NFA patients and healthy individuals. Bacteria associated with NFA demonstrate distinct communication patterns and functional implications that may influence disease pathogenesis and symptoms. Identification of specific bacterial cliques suggests potential therapeutic targets for NFA. Further research into these microbial interactions could offer novel insights into managing and treating non-food allergies. Keywords: Gut microbiota, non-food allergies, bacterial diversity, multidimensional feature selection, clique analysis

Deep-Learning Based Methods for Protein – Protein Interaction Prediction on Genomic Scale: Challenges and Opportunities

Wojciech Dec[1,2,3] Krzysztof Murzyn[1,2] Michał Markiewicz[1,2]

[1]*Department of Computational Biophysics and Bioinformatics, Jagiellonian University, Cracow, Poland*
[2]*Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Cracow, Poland*
[3]*Doctoral School of Exact and Natural Sciences, Jagiellonian University, Cracow, Poland*

Protein-Protein interaction networks (PPIN) offer a system-level insight into biological processes. However, the experimental discovery of protein-protein interactions (PPIs) on a genomic scale remains problematic, as the number of potential binary interactions that need to be considered grows quadratically with the number of proteins.

Given the slow pace of development of library-on-library techniques able to facilitate genome-wide screening of PPIs, along with other limitations, computational methods garner growing interest. Among them are deep learning based approaches, which frequently report very high accuracies; however, recently raised data leakage and reproducibility concerns put them into question. One such method proposes the use of deep hash learning (DHL) to effectively reformulate PPI prediction task from binary classification into a nearest neighbors search. This would reduce the computational complexity of predicting all-against-all PPIs in a set of proteins from $O(n^2)$ to $O(nlogn)$. Unfortunately, the effectiveness of this approach has not yet been demonstrated.

This work presents an investigation of projecting high-dimensional protein features to compact binary hash codes and their discriminative power in an all-against-all setting. Dimensional collapse is identified as the key factor limiting discriminativity, whereby the model fails to take full advantage of its capacity to encode information, resulting in embedding vectors that span a lower-dimensional subspace. Contrastive representation learning techniques and pre-trained embeddings from language models of biological sequences are employed to mitigate its effects. Further consideration is given to approximate NNS in a continuous setting and vector databases.

Mapping Mobile genetic elements in Poland's Soil Microbiome

Balakrishnan subramanian[1] Rodolfo Brizola Toscan[2] Dagmara Błaszczyk[2] Katarzyna Kopera[2] Alina Frolova[3] Paweł Łabaj[2]

[1]*Institute of Computer Science, University of Białystok*

[2]*Małopolska Centre of Biotechnology, Jagiellonian University*

[3]*Institute of Molecular Biology and Genetics of NASU, Kyiv Academic University, Kyiv, Ukraine*

Microorganisms are vital links within ecosystems and significantly contribute to One Health principles, especially due to their vast diversity in various environments, including soil ecosystems. Understanding the interconnectedness of members within these communities is paramount to comprehending their intricate influence on ecosystem health. This project explores the ecological role of mobile genetic elements (MGEs) in driving the spread of antibiotic resistance genes (ARGs) in the soil microbiomes of widely separated regions in Poland. High-quality metagenomic datasets, collected over multiple seasons and representative of different soil types and land uses, will be leveraged to identify, annotate, and characterize MGEs associated with ARGs. The objectives are to define MGE patterns and their relationships with ARG dissemination within soil microbiomes to illuminate the complexity of ARG dynamics in these environments. Additionally, the study will use environmental data, such as soil properties and climatic conditions, to explore their influence on MGEs and ARGs.

Deciphering the Role of Amino Acid Composition in Low Complexity Region Functions

Joanna Ziemska-Legięcka[1] Aleksandra Gruca[2] Marcin Grynberg[1]

[1]*Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, 02-106, Poland*
[2]*Department of Computer Networks and Systems, Silesian University of Technology, Gliwice, 44-100, Poland*

Low complexity regions (LCRs) are defined as regions in proteins that exhibit a high density of a limited number of amino acid types. LCRs can be classified as homorepeats, short tandem repeats, or simply strongly biased regions with a fuzzy amino acid composition. For a long time, these regions were considered non-functional, but further research revealed its many functions. A manual inspection of LCR sequences reveals that certain types of amino acids co-exist in these regions more often than others. This observation led us to hypothesize that the composition of amino acids in LCRs is not random and is related to their functions. To test this hypothesis, we performed statistical analyses on the dataset of LCRs from Uniref90. LCRs were identified with the SEG method using restrictive parameters. This enabled us to find mostly repetitive regions, with low numbers of false positives. In the next step, we divided LCRs into datasets. Each dataset consisted of LCRs with one type of amino acid content greater than a given threshold. We used 10 thresholds for each type of amino acid in 10% intervals between values of 10% and 90%. Based on these datasets, we calculated hypergeometric tests on two or three datasets to check if these amino acids co-occurred more frequently together than alone. As a result, we showed that some amino acids tended to co-occur in LCRs, especially similar ones in their physico-chemical properties. To examine the relationship between composition and LCR functions, we collected the most popular LCR annotations from the LCRAnnotationsDB database and compared their amino acid compositions to those of co-occurring amino acids. Finally, we hypothesize that the amino acid composition of low complexity regions is strongly connected with their functions independently of the amino acid order.

Encoding of MS images for memory management and segmentation enhancement by use of contrastive learning

Piotr Radziński[1] Maurycy Moczulski[1] Jakub Skrajny[1] Michał Ciach[2] Anna Gambin[1]

[1] *Institute of Informatics, University of Warsaw, Stefana Banacha 2, 02-097 Warsaw, Poland*
[2] *Centre for Molecular Medicine & Biobanking, University of Malta, Msida, MSD 2080, Malta*

Mass spectrometry imaging (MSI) data is known for its high resolution, which often demands substantial memory resources, making it challenging to efficiently apply analytical tools such as segmentation algorithms. This computational load complicates the detailed analysis required to distinguish different tissue states. Our research presents a novel encoding algorithm that significantly reduces the size of MSI data, easing memory constraints and enabling the use of advanced segmentation techniques. By optimizing data preprocessing through contrastive learning, our method not only simplifies the analysis process but also enhances segmentation accuracy, providing a valuable tool for precision medicine research within the MSI community.

We utilized a contrastive learning encoder-decoder architecture and carefully applied a set of loss functions to optimize the compression and preparation of MSI data for future tissue structure analysis. Contrastive loss is crucial for generating embeddings that differentiate between similar and dissimilar data points. Mean and standard loss functions help normalize the distribution of embeddings, ensuring consistent and reliable segmentation. Mean squared error loss maintains the integrity of the information within these compressed representations, ensuring no critical detail is lost. Together, these loss functions refine the data, improving segmentation by enhancing data structure and clarity for more accurate tissue state identification.

Additionally, we trained a neural network ""head"" by providing it with labeled segments of images, allowing it to extend these labels across entire MSI datasets. This significantly reduces the workload for pathologists by automating the annotation process. Our preprocessing encoding algorithm is specifically designed for MSI data to improve segmentation accuracy. It compresses large datasets into manageable sizes while preserving essential biochemical information. Initial testing on various MSI datasets and subsequent application of multiple segmentation algorithms showed improved segmentation accuracy, ranging from 4% to 34.3%, when using our encoder. These results highlight our approach's effectiveness in maintaining crucial

molecular details for precise analysis, marking a significant advancement in MSI analysis and its application in precision medicine.

Significance of Nanopore long-reads to detect methylation and build diploid genomes without parental genomic data

Sachin Gadakh[1] Karolina Jodkowska[1] Mateusz Chiliński[2] Jan Gawor[3] Dariusz Plewczynski[1]

[1] *Laboratory of Functional and Structural Genomics, Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland,*
[2] *Laboratory of Bioinformatics and Computational Genomics, Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warsaw, Poland.*
,
[3] *DNA Sequencing and Oligonucleotide Synthesis Laboratory, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warsaw, Poland*

Long-read sequencing technologies (LRS), particularly Oxford Nanopore Technology (ONT), offer several advanced applications. One notable advantage of nanopore sequencing is its capability to simultaneously detect modified nucleotides while producing long-reads, making it ideal for detecting and phasing allele-specific methylation. Recently, multiple studies have utilized long-read sequencing to generate complete or nearly complete de novo genome assemblies. Additionally, a pipeline has been developed that leverages nanopore long-reads to construct a diploid genome. This is achieved by aligning long-reads to reference genome, calling SNPs, assigning reads to haplotypes based on these SNPs, identifying breakpoints in long-read alignments, and applying the corresponding structural changes to the respective haplotypes. In our study, we analyzed samples from 3 families from the 1000 Genomes Project (1KGP), comprising 9 family members, including parents and a daughter in each family. We used the daughter's samples from these families to estimate haplotypes based on methylation and diploid assembly. Consistency between the phasing of daughters alone and trio phasing was confirmed by overlapping differentially methylated regions from the single-sample phasing and trio phasing. Furthermore, we utilized the diploid fasta sequences of the daughters' samples and their respective HiChIP reads to develop personalized HiC heatmaps and phased CTCF loops. This allowed us to study how methylation is regulated by CTCF and to determine which parent-of-origin CTCF loops influence methylation. Most of these analyses were accomplished primarily using long-reads from a single sample.

Analysis of linkage disequilibrium for next generation sequencing data.

Klaudiusz Tomczyk[1] Filip Jerzykiewicz[1] Joanna Szyda[1] Magda Mielczarek[1] Paula Dobosz[2] Magdalena Frąszczak[1]

[1] *Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw 51-631, Poland*
[2] *University Cancer Diagnostic Center, Poznan University of Medical Sciences, Poznan, Poland*

Linkage Disequilibrium (LD) can be defined as a non-random association between different loci and is commonly used in genetics. Studies based on next-generation sequencing (NGS) technology allow to identify millions of SNPs. It can be one of the reasons that analysis of NGS data has become increasingly important especially for current genetic studies. The major aim of this study is analyzing the structure of the LD coefficients between SNPs in different genomic regions and what are the factors which impact those differences. The study was performed on 41,836,187 SNPs identified in genomes of 1222 individuals of Polish origin consisted samples of 697 men and 525 women. Our analyses were mainly focused on the SNPs identified on the first and the 22st chromosomes. Among all identified SNPs in our data only 4.06% were located in genes. The research involved checking whether the LD structure significantly diers between genic (in dividing into housekeeping and other genes) region and non-genic region as well as comparing pairways LD decay across different groups based on phenotypic features. VCFtools and BEAGLE software were used to prepare initial data and calculate LD coefficients. The obtained information was divided into groups based on the maximal distances between SNPs, from 100BP up to 100.000BP. Then all statistical analysis contained Wilcoxon signed rank test, Friedmann's test, and permutation tests were performed using R and Python softwares.

Detection of copy number variants and mobile element insertions in targeted next generation sequencing data

Aleksandra Pfeifer[1] Marta Cieślicka[1] Agnieszka Pawlaczek[1] Dorota Kula[1] Jadwiga Żebracka-Gala[1] Małgorzata Kowalska[1] Anna Fiszer-Kierzkowska[1] Artur Zajkowicz[1] Magdalena Kalinowska-Herok[1] Magdalena Mazur[1] Jolanta Pamuła-Piłat[1] Karolina Tęcza[1] Mariola Szołtysik-Szot[1] Tomasz Tyszkiewicz[1] Patrycja Tudrej[1] Małgorzata Oczko-Wojciechowska[1]

[1] *Department of Clinical and Molecular Genetics, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch*

Mobile element insertions (MEIs) in coding sequences of genes are the cause of less than one percent of cases of genetic diseases, including hereditary cancers. However, MEIs are not routinely analyzed in genetic diagnostics. Copy Number Variants (CNVs) ranging from one exon to one gene are responsible for many cases of hereditary diseases. The gold standard for CNV detection is the MLPA (Multiplex Ligation-dependent Probe Amplification) method, but the next-generation sequencing can alo be used for CNV detection.

The aim of our study was to detect germline CNVs, evaluate the minimal depth of coverage required for its sensitive detection, and to detect MEIs in targeted next generation sequencing data.

We sequenced DNA from 2,700 blood samples collected from patients with hereditary cancer predisposition. We used hybridisation-capture targeted panels, and Illumina MiniSeq for next generation sequencing. Germline CNV variants were detected by ExomeDepth. The normalized depth of coverage was visualised at a single base resolution. MEIs were detected by Scramble software.

45 of 2700 samples were CNV-positive according to former MLPA. In 40 out of 45 samples, NGS also detected CNVs that were concordant with MLPA results. In one sample, quality criteria for CNV detection were not met in NGS. In four samples positive for CNV according to MLPA, NGS analysis did not detect CNV variant. In the samples, which were not previously analysed with MLPA, NGS analysis detected additional 34 CNV mutations. They were further validated by MLPA, and positive results were obtained for 19 of them.

MEIs were detected in 4 samples collected from patients with MEN1 syndrome. All the patients were members of the same family, and harboured exactly the same MEI mutation. MEI-positive patients were the same patients in which MLPA detected CNV and NGS did not. The MLPA probe is located exactly at the insertion site of the Alu element, which explains

the differences in the results of NGS and MLPA analysis.

We also performed simulations, in which we downsampled bam files to obtain 95%, 90% ... 5% of initial depth of coverage. The simulations showed that 350x depth of coverage is required to obtain the maximal CNV detection sensitivity.

Our study showed that NGS analysis can be used for sensitive CNV detection, and 350x depth of coverage is required for highest CNV detection sensitivity. NGS can also be used for MEI detection, slightly increasing diagnostic yield of NGS.

Analysis of Biological Sequences via Chaos Game Representation and Free-Alignment Methods

Adrian Kania[1]

[1] *Jagiellonian University; The Faculty of Biochemistry, Biophysics and Biotechnology; Department of Computational Biophysics and Bioinformatics*

Biological sequence analysis plays an important role in bioinformatics as it constitutes a primary step in many analyses. The traditional approach involves employing alignment for comparison, which provides information about sequence similarities and is particularly useful for tasks such as motif finding. However, alignments may not detect certain rearrangements, and their computational demands increase significantly with the number and length of sequences. An alternative set of methods includes free-alignment approaches that focus on specific features of sequences. Often, sequences are transformed into numerical representations, and various metrics are calculated. One innovative approach is based on chaos game representation, where sequences are projected onto a unit square as a set of points. These points can be interpreted as complex numbers, allowing for series analysis. Combining chaos game representation with Discrete Fourier Transform enables comparison of sequences based on their power spectra. This method proves faster and more effective compared to classical alignment approaches. While traditional chaos game representation was initially developed for nucleotide sequences, it has been extended to protein sequences as well. The effectiveness of this approach has been demonstrated on various sequence sets, such as neuraminidase genes from avian influenza viruses. Additional optimizations have been implemented to enhance computational efficiency without compromising effectiveness.

Application of Neural Network for Common Carp microbiome classification

Marek Sztuka[1] Joanna Szyda[1]

[1] *Wrocław University of Enviromental and Life Sciences*

Recently, there is a significant increase in interest in machine learning techniques and artificial intelligence. These technologies are increasingly being used in image generation, speech and text recognition, demonstrating high accuracy. They are also gaining popularity in scientific and biological fields, aiding in predictive and classification problem-solving. The aim of this study is to assess the usefulness of these tools in common carp classification based on applied probiotics. Microbiome data were collected from the intestines of carp (Cyprinus carpio). Samples were obtained from individuals living in different environments where different probiotics were added, and then sequenced. Based on this data, tables of bacterial family occurrences in individual specimens were developed. The study aimed to determine whether the type of administered probiotic could be identified based on differences in the occurrence of specific bacterial families. To achieve this, two neural network models were developed and refined: dense (DNN) and convolutional (CNN). This is a classification problem, where information about the occurrence of individual bacterial families was used as predictors. The classes represented groups of individuals fed different probiotics. Preliminary results indicate that convolutional and dense models performed similarly, with a slight advantage for DNN models. The accuracy rates were 0.6 and 0.65 for CNN and DNN models, respectively, meaning the models correctly classified 60% and 65% of cases. While these are not perfect results, further model refinement through changes in structure or training on larger datasets could significantly improve effectiveness.

playOmics: A multi-omics pipeline for interpretable predictions and biomarker discovery

Jagoda Głowacka-Walas[1,2] Kamil Sijko[2] Konrad Wojdan[2,3] Tomasz Gambin[1]

[1] *Institute of Computer Science, Warsaw University of Technology, 00-665 Warsaw, Poland*

[2] *Transition Technologies Science, 01-030 Warsaw, Poland*

[3] *Institute of Heat Engineering, Warsaw University of Technology, 00-665 Warsaw, Poland*

Background

Multi-omics analysis is increasingly popular in biomedical research. While promising, these analyses confront challenges in data integration, management, and interpretation due to their complexity, diversity, and volume. Moreover, achieving transparency, reproducibility, and repeatability in multi-omics analyses is essential for facilitating scientific collaboration and validation of complex datasets.

Results

We introduce playOmics, an open-source R package tailored for omics data analysis. It facilitates data management and biomarker discovery through various visualizations, statistics and explanations for boosted interpretability. playOmics identifies significant prognostic markers and iteratively constructs logistic regression models, identifying combinations with high predictive performance. Our tool enables users to make direct, model-driven predictions by inputting new data into the selected pre-trained model.

Conclusions

playOmics demonstrates the balance between model complexity and interpretability, crucial in biomedical research for understanding model decisions. playOmics' approach promotes a flexible model selection process, encouraging exploration and hypothesis generation in biomarker discovery. The dockerized setup and intuitive graphical interface of playOmics support its adoption in a wide range of research and clinical settings, adhering to principles of open science, enhancing reproducibility and transparency.

AI Agent-based architecture for high-throughput deep phenotyping with Large Language Models

Marek Wiewiórka[1] Wojciech Sitek[2] Tomasz Gambin[3]

[1]*Institute of Computer Science, Warsaw University of Technology, Warsaw, Warsaw 00-661, Poland*
[2]*Institute of Computer Science, Warsaw University of Technology, Warsaw, Warsaw 00-661, Poland*
[3]*Institute of Computer Science, Warsaw University of Technology, Warsaw, Warsaw 00-661, Poland*

Deep phenotyping refers to the comprehensive and detailed analysis of phenotypic traits in organisms, particularly humans, to understand complex biological processes and diseases. It is widely considered as the important component and prerequisite of precision medicine and rare diseases computationally-driven studies where detailed description of the symptoms and signs of a disease need to be encoded as standardized terms from a suitable ontology. One of the most popular ontologies for computational phenotype analysis is the Human Phenotype Ontology (HPO) that currently contains over 13,000 terms and over 156,000 annotations to hereditary diseases.

Typically, the source of phenotypic information are clinical notes that are either manually or automatically processed in the two-step procedure involving: concept recognition (finding phenotypic information in the unstructured text) and concept normalization (mapping recognized concepts to the standardized HPO identifiers). Over the years there has been developed a number of automatic methods, such as rule-based and machine learning, including recent evaluation of Large Language Models (LLMs) applicability [Yang2024, Wang2024, Gruza2024].

However, all of the above LLMs applicability studies focused only on subtasks of the automatic deep phenotyping problem: concept recognition ([Yang2024]), concept normalization ([Wang2024]) or end-to-end process but with only simple prompt engineering techniques, such as in-context learning ([Gruza2024]) not feasible for real-life systems. To the best of our knowledge there has not been proposed a production-grade LLM-based approach incorporating advanced architectures relying on Retrieval Augmented Generation (RAG) or AI agentic design patterns.

In this study we propose novel approaches and architectures for implementing a high-throughput deep-phenotyping workflow with selected commercial and open source LLMs. The proposed systems are evaluated using gold standard corpuses including BiolarkGSC+ and ID-68.

LLM-based Approach for Extracting Genomic Region Coordinates from Biomedical Publications

Wojciech Sitek[1] Maria Bochenek[2] Marek Wiewiórka[1] Anna Gambin[2] Tomasz Gambin[1]

[1]*Institute of Computer Science, Warsaw University of Technology*

[2]*Institute of Informatics, Warsaw University*

Biomedical publications in PubMed often contain unstructured data, making it difficult to reproduce experiments and reuse presented information in future research. Creating datasets and benchmarks based on genomic sequence data, clinically combined with disease, variants and other metadata, often requires manual information retrieval from certain publications. This issue is particularly evident with genomic region coordinates, which are rarely clearly stated and structured in a predictable way within the text. We propose a semi-automatic solution that leverages Large Language Models (LLMs) to extract relevant genomic region coordinates and their immediate text context from selected PubMed publications. This retrieval is performed in the context of the given user query, that may contain: genes of interest, types of genomic regions (e.g. variants, functional regions), and associated disease. A sample query [gene="POU3F4" ; disease="X-linked deafness type 2"; types-of-regions="variants"; publications="de Kok et al. 1996; Ahn et al. 2009; Naranjo et al. 2010"] will search for genomic coordinates of variants associated with the X-linked deafness type 2 within the gene POU3F4 or its regulatory regions. Our tool fetches full texts of selected publications and their references from PubMed Central. Next, we use LLM to extract and normalize (by performing conversion to HGVS format) all of the occurrences of genomic region coordinates. Another module assesses the relevance of obtained results to the given user query. If within the text there are any ClinVar, dbSNP or other external identifiers available, they are used to obtain relevant coordinates. We evaluated the usability of various LLMs, including highly-ranked closed and open-source models by comparing their F1 performance using a manually curated truthset. The proposed solution aims to streamline the creation of new datasets and benchmarks based on genomic sequences.

Rapid and Accurate Estimation of Genetic Relatedness Between Millions of Viral Genome Pairs Using MANIAC

Wanangwa Ndovie[1,2] Jan Havranek[3] Janusz Koszucki[1,2] Jade Leconte[1] Leonid Chindelevitch[4] Evelien Adriaenssens[5] Rafal Mostowy[1]

[1] *Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland*

[2] *Doctoral School of Exact and Natural Sciences, Jagiellonian University, Krakow, Poland*

[3] *The Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Krakow, Poland*

[4] *The Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, UK*

[5] *Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom*

Average Nucleotide Identity (ANI) is a metric that is used for species delineation and quantifying the similarity between microbial genomes. While ANI was originally developed with bacterial genomes in mind, in recent years it has become commonly used to assess relationships between viral genomes, including bacteriophages. There are several tools currently available for systematic calculation of ANI between pairs of genomes such as Pyani, fastANI and Mash. However, these tools have not been systematically assessed on viral samples and do not take into account the high degree of genetic mosaicism between some viruses. Furthermore, it is unclear how well the current tools scale to large datasets. Here we present a new approach called MMseqs average nucleotide identity calculator (MANIAC) that uses MMseqs2 to systematically calculate ANI and alignment fraction (AF) values for pairs of genomes by applying the same approach used by Goris and colleagues. We assessed the accuracy and speed of MANIAC using 500 NCBI refseq phage sequences and found that it reproduces the accuracy of Pyani (R2=0.999), a BLAST-based tool, at about 1/100 of running time. We then further assessed the performance of MANIAC by generating an in-silico dataset of 6840 simulated pairs of DNA sequences with a known ANI and AF values. Our results showed that MANIAC, with the right parameterization, was able to accurately predict ANI values even below the Pyani prediction threshold of 60% ANI. We were able to process the entire NCBI refseq (4500 phage genomes) in 3 hours and the Inphared database (24000 phage genomes) in approximately 36 hours, making MANIAC applicable for intermediate-to-large datasets. Altogether, MANIAC is a fast, accurate and scalable tool for the analysis and investigating genetic relationships between pairs of viral genomes.

RNA secondary structure modeling using loops decomposition and integer programming

Olga Karelkina[1]

[1] *Systems Research Institute, PAS*

RNA secondary structure prediction methods often involve estimating the free energy of all possible structures for a given RNA sequence. To determine this free energy, a secondary structure can be uniquely decomposed into characteristic substructures known as loops. The energy of each loop is calculated using empirically derived nearest neighbor rules and associated parameters, with the total free energy of the structure being the sum of the energies of its constituent loops.

The proposed integer programming model employs free energy minimization approach to find the optimal secondary structure for a given RNA sequence, considering all possible loop decompositions. The model defines the search space of feasible structures and incorporates different loop types through a set of logical constraints. Additionally, an extended parameterized version of the model allows for obtaining suboptimal structures and structures with specific characteristics.

The accuracy of the loop decomposition-based integer programming model has been benchmarked against RNA sequences of small and medium sizes with known structures from the ARCHIVE II dataset. Preliminary computational results demonstrate that the model generates secondary structures with base pairs and energy values that are close to or match those of the reference structures.

hadexversum: HDX-MS analysis made easy

Weronika Puchała[1] Michał Kistowski[1] Michał Dadlez[1] Michał Burdukiewicz[2,3]

[1] *Institute of Biochemistry and Biophysics, Polish Academy of Sciences*

[2] *Clinical Research Centre, Medical University of Białystok*

[3] *Vilnius University, Institute of Biotechnology*

Hydrogen deuterium exchange (HDX) is a unique technique of protein structural studies as it provides insight into the dynamic properties of protein chains entangled in more or less stable spatial structures. The HDX-MS data is complex and it requires a dedicated solution. To answer the community's needs, we developed a family of applications - hadexversum, personalizable workflows for multidimensional challenges.

The very first created tool is HaDeX - a versatile software for processing, analyzing, and visualizing output data on the peptide level. The application provides a complete analytic workflow, with precise uncertainty calculations and reporting features recommended in the community guidelines. Various visualization methods (including novel ones) ensure in-depth data exploration.

To expand the possible overview of the exchange phenomenon, we created HRaDeX. We utilized a well-known mathematical exchange model and proposed a unique method of uptake curve classification, using color code corresponding to the exchange pattern. Ultimately, the classification results are aggregated to the high-resolution level and presented on the 3D structure, enabling the correspondence of the spatial data with experimental results. As HRaDeX works on one biological state at a time, compaHRaDeX provides comparative analysis to ensure comfort of use.

All applications are available as a web-servers, R-packages, and standalone software. The excessive documentation accompanies them to ensure maximal transparency as an open-source solution (https://github.com/hadexversum/).

Day 2 - 12 September 2024

Keynote speaker

Fantastic world of tandem repeats and how to characterize them

Fritz Sedlazeck[1]

[1] *Baylor College of Medicine*

Tandem repeats are hard to assess but important regions on the human genome. There are around 200 pathogenic and phenotypic tandem repeats, which are often ignored by standard analysis. Furthermore, these tandem repeats are utilized in crime scene investigations. Given the impact of these regions, it is important that we learn more about these regions of the human genome and their role in diseases and other traits.

Session 3

Quantitative analysis of tRNA modifications by nanopore RNA sequencing

Wiktor Kusmirek[3] Natalia Strozynska[1] Paula Martin-Arroyo[1] Katarzyna Pietka[1] Grazyna Leszczynska[2] Robert Nowak[3] Malgorzata Adamczyk[1]

[1] *Warsaw University of Technology, Faculty of Chemistry, Laboratory of Systems and Synthetic Biology, Warsaw, Poland*
[2] *Lodz University of Technology, Faculty of Chemistry, Institute of Organic Chemistry, Lodz, Poland*
[3] *Warsaw University of Technology, Faculty of Electronics and Information Technology, The Institute of Computer Science Warsaw, Poland*

Nanopore sequencing has been proven a reliable method for direct RNA sequencing. The method allows for evaluation of chemical modification present on mRNA. tRNA modifications are typically identified and quantified with high accuracy using LC-MS methodologies or NGS-based sequencing, which despite recent improvements suffer from technical pitfalls. New efficient tRNA sequencing methods are still needed for major discoveries in epigenetics. We have developed a new methodology called MODetRNAseq, a nanopore-based approach to sequence native tRNA molecules and to identify modifications at uridine 34 (U34) position. We designed synthetic, modified oligonucleotides for full tRNAs assembling and used them as internal standards in ONT sequencing to enhance processing efficiency of raw nanopore current intensity signals obtained for natural tRNAs. We developed conditions for the oligonucleotides splint ligation with singly modified ncm5U-RNA and ncm5S2U-RNA. An artificial neuronal network external model has been trained on raw sequencing data and was utilised to identify specific signals containing elements characteristic to the modifications in different tRNAs. The basecalling process was conducted on the fast5 raw reads utilizing the Guppy and Dorado applications, the Dorado basecaller was launched in two modes: fast and high-accuracy, resulting in three separate datasets. Analysis of these datasets demonstrated that the selection of basecaller and model significantly influences the detection of tRNA modifications. Furthermore, we established that the mapping parameters of tRNA reads to reference sequences are critical in the modification detection process. Our study also revealed that tRNA modifications can be identified at the level of current value changes, without the basecalling step.

Funded by POB Biotechnology and Biomedical Engineering, Warsaw University of Technology Research University (IDUB) BIOTECHMED-3 Advanced No. 504/04496/1020/45.010421 (MA).

A comprehensive pipeline for gene expression and mutation profiling via targeted sequencing with unique molecular barcodes

Michal Marczyk[1,2] Chunxiao Fu[3] Lili Du[3] William Fraser Symmans[3]

[1]*Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*
[2]*Yale Cancer Center, Yale School of Medicine, New Haven, CT, USA*
[3]*Departments of Pathology and Translational Molecular Pathology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA*

Targeted sequencing provides increased sensitivity, dynamic range, reduced cost, and increased throughput compared to standard sequencing of the whole genome or transcriptome. It is also advancing as an alternative translational research and clinical testing technology to RT-PCR or oligonucleotide microarrays. However, specific data processing and analysis methods are needed since only selected transcripts are measured. Using a spike-in control dataset with different percentages of transcripts with specific mutations in the ESR1 gene (2.5%, 5%,10%, and 25%), we developed a pipeline for gene expression quantification and mutation profiling for targeted sequencing data with unique molecular barcodes. The pipeline works for any number of gene primers and includes the following steps: (i) quality control and adapter trimming with Trimmomatic; (ii) searching for primer sequences with mismatches and extraction of UMIs; (iii) sequence alignment to transcriptome using STAR; (iv) filtering for selected genes; (v) UMI directional clustering; (vi) estimating consensus read for each unique UMI; (vii) gene quantification and variant allele fraction estimation. Using Trimmomatic led to an increased mean base quality score in all samples. Allowing for mismatching during primer matching increased the number of single matches and single-primer reads. Most reads were filtered during these first two steps. Almost all UMIs were 8bp long as expected. More than 80% of reads were uniquely aligned to transcriptome with a low multimapping rate. Filtering by tag removed an additional 1% of reads. Clustering UMIs reduced the number of false positives. Normalized gene expression was similar between samples. Finally, a correct increase in allele fraction of ESR1 gene mutations was observed in subsequent samples. By analyzing spike-in data we showed that the proposed pipeline can correctly estimate expression of selected genes and allele fraction of single nucleotide variants. This research was funded in part by the National Science Centre, Poland grant no. 2023/50/E/NZ2/00583 (MM).

Enhancing RNA-seq Fusion Detection: A Meta-Analysis of Benchmark Variability and Tool Performance.

Iga Ostrowska[1] Tomasz Gambin[1]

[1] *Warsaw University of Technology*

Many tools have been developed for RNA-seq fusion analysis, such as Arriba, STAR-Fusion, and FusionCatcher. However, no single tool can be considered universal. Our study compared the performance of 32 tools (specific versions not included) across various types of transcriptomic data from 10 benchmarks, revealing significant performance variability. We conducted a meta-analysis of fusion detection methods, considering various tools and datasets with different characteristics. This comprehensive analysis evaluated multiple dimensions of performance, including precision, recall, and F-measure. We also assessed the reproducibility of results across benchmarks by analyzing the variance in reported performance for each tool and dataset combination. To facilitate this meta-analysis, we developed a custom pipeline to extract relevant performance statistics from source publications. This pipeline allows for easy updates of the meta-analysis with new publications. Additionally, we partially verified the collected results using the nf-core/rnafusion pipeline, which integrates three popular fusion analysis tools. From the analysis of 10 selected benchmarks, testing 32 different tools and 31 datasets (both unique and repeated across benchmarks), we identified key factors influencing the performance of fusion detection tools. Based on these findings, we proposed a recommendation system for selecting appropriate tools depending on the type of data analyzed. We also observed significant variability in result presentations across benchmarks, affecting both the ease of information extraction for meta-analysis and the reproducibility and confidence in the results. One of our key findings indicates that benchmarks should include data from both state-of-the-art and cutting-edge technologies (e.g., whole transcriptome analysis) as well as data commonly used in clinical practice (e.g., gene panels). This approach ensures comprehensive evaluation and practical applicability of fusion detection tools. Finally, our recommendation system can also be used to improve the performance of ensemble approaches such as MetaFusion. By incorporating data characteristics into the ensemble method, our system can refine the selection of tools and enhance the reliability of fusion detection outcomes, moving beyond the simplistic majority voting approach.

Session 4

A three-level modeling for identifying important predictor variables in genome-wide association studies suffering from p » n.

Jakub Liu[1,2] Dawid Słomian[3] Paula Dobosz[2] Joanna Szyda[1,2]

[1]*Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland*
[2]*University Cancer Diagnostic Center, Poznan University of Medical Sciences, Poznan, Poland*
[3]*National Institute of Animal Breeding, Cracow, Poland*

Background/Objectives Many genomic datasets are characterized by a much greater number of explanatory variables (p) than the number of observations (n). As a consequence, it is not possible to fit all the variables into a single statistical model.

Methods A three-level approach was applied to the association analysis of COVID-19 susceptibility with SNPs identified in whole genome sequences of 1222 individuals. The first stage consisted of fitting k logistic regression models with subsets of SNPs that guarantees $p < n$. Each SNP was present in at least one of the k models. In the second stage, goodness-of-fit of each model was used as a dependent variable in a single Elastic Net Regression model that incorporated effects of all explanatory variables estimated in the first stage. In the third step, the effects of the p explanatory variables were clustered using the k-means clustering algorithm to identify SNPs that are important predictors of the susceptibility to COVID-19. Results For the original set of 42 million SNPs, two groups representing SNPs important and irrelevant to being resistant or susceptible to COVID-19 infection were identified.

Conclusions This approach enables the selection of explanatory variables for the situation of $p < n$, for which fitting all variables in the same model would enforce shrinkage while fitting a multitude of single SNP models poses problems with multiple testing. Even though the method is computationally intensive, it is well suited for parallelization and, hence, computationally feasible.

Efficient and Accurate LC-MS Feature Alignment Using Sliced-Wasserstein Distance.

Justyna Król[1] Anna Gambin[1] Barbara Domżał[1] Błażej Miasojedow[1]

[1] *University of Warsaw*

Liquid Chromatography-Mass Spectrometry (LC-MS) is a powerful analytical technique widely used in diverse fields such as proteomics, metabolomics, and environmental science for identifying and quantifying complex mixtures. A critical challenge in LC-MS data analysis is feature alignment, which involves matching features (peaks) across multiple chromatograms to ensure accurate and consistent compound identification. This process is complicated by retention time shifts, variations in peak intensity, and noise. Optimal transport theory has been previously employed in mass spectrometry to address various challenges, including LC-MS feature alignment. While the Wasserstein distance is a potent tool for handling one-dimensional mass spectra, it suffers from high computational complexity when applied to LC-MS data. We propose a novel method that leverages the sliced-Wasserstein distance for feature alignment, significantly reducing time complexity while maintaining alignment accuracy. The sliced-Wasserstein distance is a variant of the Wasserstein metric that offers a computationally efficient way to measure discrepancies between distributions, making it well-suited for high-dimensional LC-MS data. Additionally, the proposed algorithm incorporates noise reduction techniques from optimal transport applications in one-dimensional mass spectra analysis, further enhancing its robustness and precision in feature alignment tasks.

Unsupervised learning for detecting relative importance of pathway activity in individual cells based on scRNA-Seq data

Anna Mrukwa[1] Joanna Zyla[1]

[1]*Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland*

Pathway enrichment is one of the main steps in bioinformatical analysis. With the development of scRNA-Seq, the single-sample PE algorithms became more popular as they offer the way to access the singular gene expressions in pathways, giving the ability to precisely investigate each deviation across cells. Yet, the challenge of such analysis is to find the relative importance of pathway activity in singular cell. Here, we propose unsupervised pipeline to group pathway activity across investigated cells. Four scRNA-seq sets of known immunological cells and different volume, were collected. Gene expressions were log-normalized and transformed into pathway activity scores (PAS) using sets of genes representing signatures of immunological processes. Gene sets were extracted from publicly available databases and marked into target cell type. Transformation into PAS was performed by the six known single-sample algorithms (CERNO, AUCell, PLAGE, ssGSEA, Z-score and Mean). Then, we propose usage of Gaussian Mixture Modeling with BIC criterion on each PAS vector. To extract relatively important pathway activity in each cell, the component with the highest mean was taken. To assess results ARI and FDR were calculated based on initial cell types and pathway target. This approach was compared to the sole existing solution based on heuristic statistic: AUCell thresholding. Our GMM based method outperforms AUCell thresholding in terms of FDR for each tested PAS transformation. Also, ARI shows that proposed method is better in AUCell, CERNO and ssGSEA algorithms. Investigation in separate cell types shows that for B cells results remain unchanged. Yet, for hardly separable NKs there is no difference between methods. To sum up, our method detects relative importance of pathway activity across investigated cells for data of different volume and variety PAS transformation. Moreover, the proposed solution is fully unsupervised in contrary to the existing solution in the field.

Deep into the Dark Proteome: structural disorder analysis of low-complexity subtypes in protein sequences using the AlphaFold pLDDT metric.

Barbara Ilnicka[1] Sylwia Szymańska[2] Aleksandra Gruca[2]

[1]*Faculty of Automatic Control, Electronics, Information Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice*
[2]*Departament of Computer Networks and Systems. Silesian University of Technology, Gliwice, Poland*

Low-complexity regions (LCRs) are fragments of protein sequences characterized by low amino acid diversity. In many studies, LCR regions have been linked to the disordered structure of proteins, but a comprehensive study of the relationship between LCRs and protein structure has always been a challenge, because protein structure databases mainly included globular proteins. Recently, researchers have become interested in the Dark Proteome, which is defined as intrinsically disordered proteins or proteins containing intrinsically disordered regions. Despite the increasing understanding of LCRs in proteins, they continue to be ambiguous elements of the proteome. Advances in deep neural networks, resulting in development of AlphaFold, have greatly improved our ability to analyze protein structures which may help in our understanding the structural consequences of LCRs presence. The definition of LCRs varies depending on the distinct characteristic features like structure, amino acid composition or its periodicity, which allows us to divide them into subtypes. The subtypes including homorepeats, short tandem repeats, fused regions, compositionally biased regions, intrinsically disordered regions, prions, and amyloids were identified using specialized tools and databases like GBSC, SEG, CAST, IUPred3, PrionScan and AmyloGraph database. Then we analyzed their structural differences using predicted Local Distance Difference Test (pLDDT) values. The pLDDT metric, a confidence indicator for AlphaFold structure predictions, also serves as a predictor for structural disorder. Our research highlighted the non-globular and disordered structural nature of nearly all LCR groups, indicated by median pLDDT values falling below the threshold of 50. Only amyloids and LCR identified with SEG default exhibited a median pLDDT value above 80. Despite variations in the lengths of the analyzed regions, the pLDDT values remained consistent across most LCR subtypes.

This study reinforced our knowledge of the structural properties of different types of LCRs and their propensity for disorder. By investigating the overlap between fragments belonging to different subgroups, we showed that LCRs frequently intertwine and often present characteristics of multi-

ple subtypes simultaneously. The low median pLDDT values obtained for almost all analyzed subtypes confirm strong relation between existence of LCRs and protein structural disorder.

An automated analysis of homocoupling defects using MALDI-MS and open-source computer software

Maria Bochenek[1] Michał Aleksander Ciach[1,2,3] Sander Smeets[4,5,6] Omar Beckers[4,5,6] Jochen Vanderspikken[4,5,6] Błażej Miasojedow[1] Barbara Domżał[1] Dirk Valkenborg[2] Wouter Maes[4,5,6] Anna Gambin[1]

[1]*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Banacha 2, Warsaw, 02-097, Poland*
[2]*Data Science Institute, Hasselt University, Hasselt, 3500, Belgium*
[3]*Department of Applied Biomedical Science, Faculty of Health Sciences, University of Malta, Msida, MSD 2080, Malta*
[4]*Institute for Materials Research (IMO), Hasselt University, Agoralaan, Diepenbeek, 3590, Belgium*
[5]*IMEC, Associated lab IMOMEC, Wetenschapspark 1, Diepenbeek, 3590, Belgium*
[6]*Energyville, Thorpark, Genk, 3600, Belgium*

Push-pull conjugated polymers are organic semiconductors that can be potentially used in organic solar cells and photodetectors, however, due to polymerization reaction characteristics some polymers' chains may contain structural defects such as homocoupling of identical building blocks or seemingly unexpected end-groups. To detect the presence of the homocoupling MALDI-ToF mass spectrometry can be used, however, it is important to note that this method does not provide detailed knowledge about polymer structure.

One of the ways to estimate the quantities of analytes in a mass spectrum is using Wasserstein regression. In practice, annotating the polymer spectra requires large reference compound spectra libraries, which can lead to false positives since more complex models tend to provide a better fit for the experimental data. Additionally, based on expert knowledge, it is possible to gauge which compounds are more likely to be present than others.

This paper presents Masserstein package extension that decreases the false positive rate of annotations with extensive libraries of spectra by allowing users to specify a penalty for annotation, either LASSO-style where adding any additional annotation inflicts the additional cost which in turn promotes annotations including fewer compounds or by specifying penalties for annotation for each compound individually, therefore incorporating prior knowledge about which compounds are more likely to occur. We provide a rigorous description of the optimization problem incorporating the penalties for annotations and show how it can be converted to a linear program. The algorithm for solving this optimization problem using the Simplex method has been implemented and added to the open-source `masserstein` package.

We used the quantitative information about estimated polymer concen-

trations, provided by Massesrstein, to quantify homocoupling defects using the proposed formal homocoupling measure, monomer frequency, and end-group distribution. We found that not only Masserstein annotates the spectra as well as human experts but those annotations preserve the repeating end-groups pattern between groupings of polymers sharing identical monomers' counts. In conclusion, the proposed pipeline provides rapid screening for homocoupling defects and can be used to compare the amount of homocoupling between different samples.

Sponsor

Leveraging genomics at scale for Drug Discovery: the AstraZeneca Genomics Initiative

Sebastian Wasilewski[1]

[1] *Centre for Genomics Research, Discovery Sciences, BioPharmaceuticals R&D, AstraZeneca, Cambridge, UK*

The discovery and development of new drugs is a complex and challenging endeavour, particularly when considering the genetic diversity inherent in human populations. This talk will focus on the critical role that rare variant genetic plays in the drug discovery process. Unlike common variants, highly penetrant rare genetic variants can provide unique and actionable insights into disease mechanisms and therapeutic targets. However, identifying and understanding rare variants necessitate comprehensive population scale sequencing and data analysis efforts. To illustrate the impact of such large-scale genetic investigations, we will delve into the AstraZeneca Genomics Initiative, an example of how industry leadership can drive forward genetic research. This initiative showcases the essential nature of public-private consortia and partnerships in producing genetic and omics data at a scale required for impactful drug discovery. Furthermore, the talk will highlight the importance of broadening the genetic diversity to include understudied populations and geographies. The landscape of genomic data is historically skewed towards certain ancestries, by expanding the diversity of R&D data we can work towards treatments that are more efficacious across a diverse global population.

Day 3 - 13 September 2024

Session 5 - PTBI Laureates

Computational methods for anti-cancer drug sensitivity prediction

Krzysztof Koras[1]

[1] *University of Warsaw, Ardigen*

Computational models for drug sensitivity prediction have the potential to significantly improve personalized cancer medicine. Drug sensitivity assays, combined with molecular profiling of cancer cell lines and drugs become increasingly available for training such models. Existing models largely differ in terms of the modeling framework, utilized data and modeling objectives. This thesis is devoted to comprehensive modeling of drug sensitivity data and builds upon three projects. In the first one, we comprehensively developed and evaluated several feature selection strategies for per-drug sensitivity prediction. In the second, we developed a deep recommender system for prediction of kinase inhibitors efficacy across cancer cell lines, with a tailored model interpretability approach. The third project established a methodology for clustering of the latent data representations within a variational autoencoder framework, with an application to drug sensitivity prediction and new compounds generation. The thesis highlights crucial challenges regarding the problem of drug sensitivity prediction problem and provides several means to address them. Specifically, research topics include feature selection, multi-task learning, model intepretability, representation learning and generative modeling. The research presented in the thesis naturally evolved from using well-established machine learning models with more emphasis put on data exploratory side, to developing custom methods based on neural networks and generative modeling, introducing novel technical solutions.

Modeling and simulation of multi-agent systems representing processes in the RNA World hypothesis

Jarosław Synak[1]

[1] *Poznan University of Technology*

The origin of life on Earth is one of the most interesting problems in science. To this day it remains a mystery how from inorganic matter something as complex as cells could have emerged. One of the biggest obstacles is the lack of paleontological data from the period, which means all the theories have to be based on modern day wet experiments and theoretical analysis. The goal of my thesis was to focus on the latter and explore various methods, based on mainly on computer simulations, which can be used to tackle the problem. Several different models were discussed, each with its own set of assumptions and laws governing its evolution. All of them were also analysed mathematically using differential equations and similar methods. This theoretical analysis was crucial to discover general rules which determined the survival of populations of self-replicating RNA molecules. It is worth mentioning that the variety of models allowed to analyse the problem from different angles, with multi-agent models focusing on single RNA molecules, while others treated the system as a chemical solution or placed emphasis on interactions between different populations of RNAs.

Robinson-Foulds distance between phylogenetic networks and gene trees

Natalia Rutecka[1]

[1] *University of Warsaw*

Abstract: Phylogenetic networks enable the modelling of both vertical and reticulate evolution of species, including hybridisation, recombination, and horizontal gene transfer. Due to the vast space of phylogenetic networks, it is often constrained to smaller subclasses, eg. relaxed tree-child networks. A fundamental concept related to phylogenetic networks is that of displayed trees, i.e., trees derived from a network by removing a set of reticulation edges. > This dissertation addresses the NP-hard problem of computing the Robinson-Foulds cost between a gene tree G and an optimal tree displayed by a given relaxed tree-child network N. The computation of the cost can be used to compare alternative hypotheses regarding the reticulate evolution of a group of species, or to infer their phylogenetic network from a collection of gene trees. We propose an algorithm for the problem with a complexity of $O(2^r \cdot |G|^2|N|)$, where r denotes the number of reticulation events in N. Additionally, we state a hypothesis that the complexity can be improved to $O(2^r \cdot |G||N|)$. > Our simulation study demonstrates that the average complexity of the algorithm is significantly lower than the pessimistic complexity, indicated by the exponential factor of 2^r being reduced to between $2^{0.33r}$ and $2^{0.55r}$. This underscores the potential of our algorithm to process instances with 2-3 times more reticulations than the naive approach. Additionally, the utility of our algorithm is demonstrated through an experiment in which we infer a phylogenetic network of 15 coronavirus species.

Algorithm for constructing a variation graph from a colored de Bruijn graph

Adam Cicherski[1]

[1] *University of Warsaw*

Reference genomes are crucial genetic resources, serving as coordinates for gene annotations and read mapping targets, enabling downstream analysis. However, reliance on a single reference genome introduces bias towards reference alleles, especially in populations that are diverse or genetically distant from the reference. Furthermore, some genetic variants are challenging to characterize relative to a reference genome. Due to these issues, there is a growing preference for *pangenome* graph models to represent multiple references simultaneously.

Variation graphs (VGs) and de Bruijn graphs (dBGs) are prominent in this regard. dBGs consist of nodes labeled with k-mers and edges representing $k-1$ overlaps, allowing for straightforward construction in linear time. VGs, on the other hand, use nodes labeled with DNA sequences of arbitrary length, representing genomes via concatenation of labels without the redundancy of dBGs. VGs provide a clear interpretation and coordinate system for annotations, but their construction is more computationally intensive. Moreover, their flexible definition leads to many possible VGs for the same sequences, with existing construction criteria being mainly heuristic. Hence, our focus was on establishing strict criteria for VG construction, particularly exploring whether the relationships between residues in dBGs could be translated into VG frameworks.

We introduced two attributes of graph representation: *k-completeness* and *k-faithfulness*, both of which are trivially satisfied by dBGs but are essential for defining the desired VG.

Let G be a graph equipped with a set of genomic paths π representing a sequences set S. In simple terms, the representation is k-complete if every common k-mer in S is depicted by the same path, and it is k-faithful if all multiple occurrences of a vertex in π are essential to satisfy k-completeness. To be more precise, multiple occurrences are allowed only if:

- They are encompassed within a shared subpath labeled with a string of length $\geq k$.
- They arise as a consequence of the transitive closure of this relation.

We demonstrated that a VG meeting these criteria exists for any genomes collection, using an algorithm that preserves attributes during the dBG to

VG transformation. Additionally, this VG is unique up to transformations involving the splitting or merging of vertices on unbranching paths.

Determining gene expression profile in pituitary somatotroph tumors and identification of their molecular subtypes

Julia Rymuza[1,2] Mateusz Bujko[2] Monika J. Piotrowska[1]

[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*
[2] *Department of Molecular and Translational Oncology, Maria Sklodowska-Curie National Research Institute of Oncology*

Acromegaly is a serious and life-threatening disease. Most often (95% of cases) it is caused by excessive secretion of growth hormone by Pituitary Neuroendocrine Tumors (PitNET). The clinical features of *acromegaly* may be subtle or severe and vary from limb enlargement, soft tissue swelling, joint pain and hyperhidrosis to frontal bone hypertrophy, diabetes, hypertension, respiratory and cardiac failure.

The aim of this study was to determine the gene expression profile of somatotroph PitNETs and attempt to identify the molecular subtypes of these tumors. The study used RNA sequencing data from 48 somatotroph tumor samples. The analysis was performed using well-established bioinformatics tools such as `HISAT2`, `featureCounts` and `DESeq2`. Molecular subtypes of tumors were distinguished using k–means algorithm. They were then characterized using differential gene expression analysis. Additionally, the importance of genes with differential expression levels in molecular subtypes was investigated using functional analysis.

Understanding the molecular mechanisms underlying the pathogenesis of pituitary tumors is essential for improving the standards of patient care and individualizing the therapeutic process. Presented results may be helpful in developing a preliminary molecular subclassification of tumors, which, according to the current criteria, are collectively recognized as somatotroph PitNETs. Additionally, in the future, the obtained results may enable the identification of potential molecular biomarkers characterizing the molecular subtypes of somatotroph tumors.

Creation and evaluation of an amino acid substitution matrix for low complexity fragments of proteins combined with the improvement of a method to compare these fragments

Maciej Dzikowski[1]

[1] *Faculty of Mathematics, Informatics, and Mechanics, Univeristy of Warsaw 1*

Nowadays, the interest of the scientific community has shifted to include parts of the proteome that were previously deemed as useless in the sequencing process. This revealed a new purpose of some sequences with a biased composition, as sequence analysis is an elementary approach to hypothesise about its functions and possible applications. The most common methods in this field utilise substitution matrices as a fundamental part of their algorithms. By default, those techniques employ matrices suitable for standard (high complexity) sequences. When analysing data sets with a biased composition, such as low complexity regions (LCRs), it would be best to use matrices designed for these purposes. LCRs are an example of a specific type of short sequences with a low diversity in their amino acid composition. Along with the constantly increasing interest in these regions, the demand for new methods and tools created specifically for them also rises, especially due to a long period of omitting these fragments in research. Therefore, this project addresses this need by proposing a newly constructed substitution matrix dedicated to low complexity regions. This thesis presents a novel, comprehensive procedure for constructing substitution matrices from a collection of protein sequences with assigned family membership. It involves the identification of proteins with the same origin, the preparation of alignments, and the utilisation of a slightly improved method that was developed for the BLOSUM series of matrices. This procedure is followed by an evaluation that confirms the quality and purpose of the outcome obtained. With the LCR-matrix, existing sequence analyses programs can be easily adapted for low complexity regions. The presented approach is fully reproducible and can serve as a basis for the construction of other matrices for various proteins with unusual properties.

Evaluation of the Efficiency of Signaling Pathway Activation Scores Matrices Using Unsupervised Machine Learning Techniques in scRNA-Seq Data

Kamila Szumala[1] Joanna Zyla[1]

[1] *Silesian University of Technology, Department of Data Science and Engineering, Gliwice, Poland*

Next-generation sequencing (NGS), particularly single-cell RNA sequencing (scRNA-seq), enables comprehensive gene expression profiling at the single-cell level, producing high-dimensional data that pose significant analytical challenges. To explain cellular heterogeneity, dimensionality reduction techniques and pathway activation score (PAS) matrices are applied. These approaches aggregate gene expression data into biologically interpretable sets representing cellular processes or signaling pathways.

This study evaluates various algorithms for transforming gene expression matrices into PAS matrices, including ssGSEA, AUCell, and PLAGE. The performance of these algorithms was assessed based on their ability to separate cell groups according to expert-labeled references and the quality of their clustering outputs using four popular unsupervised machine learning algorithms, including hierarchical clustering and Louvain clustering. The Silhouette Index (SI) and Davies-Bouldin Index (DBI) were used to assess the quality of transformation and the level of retained information about cell groups, while the evaluation of unsupervised algorithms was based on ARI, FMI, and NMI metrics.

Both SI and DBI showed that the worst ability to separate cell types is achieved by a matrix consisting of the expression level of the most differentially expressed genes (SI = 0.03, DBI = 4.8). Of all the algorithms tested, the PLAGE (SI = 0.18, DBI = 1.76) and SparsePCA (SI = 0.19, DBI = 1.78) algorithms achieve the highest results. For the quality of clustering relevance, SparsePCA (ARI = 0.52, NMI = 0.7, FMI = 0.63) and PLAGE (ARI = 0.48, NMI = 0.68, FMI = 0.59) methods show the same or slightly higher values as the standard gene-level analysis method (ARI = 0.5, NMI = 0.68, FMI = 0.59).

The results indicate the effectiveness of using transformation algorithms in scRNA-seq data analysis. The study found that choosing the right PAS transformation algorithm significantly affects the interpretability and biological relevance of scRNA-seq data analyses. Moreover, the most effective algorithms for this purpose were identified as PLAGE and SparsePCA. These findings underscore the importance of method selection in reducing dimensionality while preserving key biological insights, facilitating a deeper understanding of cellular heterogeneity and complex biological processes.

Keywords: scRNA-seq, unsupervised learning, Pathway Activity Score

Quality Assessment of 3D RNA Structures Using Graph Neural Networks

Bartosz Adamczyk[1] Maciej Antczak[1] Marta Szachniuk[1]

[1] *Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology*

Reliable quality assessment and ranking of 3D RNA structures are crucial to identifying limitations and practical applications of their models designed in silico. Currently, the accuracy of computational predictions is mainly analyzed based on the reference structure, but we increasingly need to evaluate 3D RNA models when no reference is known. To face this issue, we applied graph neural networks (GNNs) in modeling inter-residue relationships and evaluating the quality of 3D RNA structures. The developed GNN-based system operates on local descriptors of the RNA structure. It was trained on a diverse set TS of non-redundant data sourced from 737 experimentally determined RNAs. To construct the training set, we extracted descriptors related to each ribonucleotide of each RNA in the TS collection. The same was done for nucleotides in the set of 17,790 RNA models predicted by RNAComposer for every target in TS. The 870,367 descriptors derived from the predicted models were scored in terms of their compatibility with the corresponding target descriptors. This provided us with RMSD values that together with the respective descriptors could fuel the GNN. The system was benchmarked against ARES, currently the best tool for reference-free evaluation of 3D RNA models, and tested on a set of RNA-Puzzles predictions. Both experiments proved the GNN efficiency in evaluating and ranking 3D RNA models in terms of their quality. We hope that the presented method will contribute to new standards for the assessment of 3D RNA structures and serve to filter untrustworthy conformations produced by RNA prediction methods.

Keynote speaker

Modeling spatial omics profiles of tumor microenvironments

Ewa Szczurek[1,2]

[1]*Institute of AI for Health at Helmholtz Munich, Germany*

[2]*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*

Recent development of spatial omics techniques enabled high resolution profiling of the transcriptome and proteome of the tumor microenvironment. However, the spatial data poses several challenges, from the identification of cell types and cander clones, through discovery of spatial niches, to the determination of tumor infiltration patterns. In my talk, I will present methods developed in my lab that tackle these challenges. We apply our methods to data from prostate, breast and lung cancer samples, identifying interesting spatial signals in those cancers.

Sponsors


aws


AstraZeneca


bio forum
Central European


POLITECHNIKA WARSZAWSKA
WARSAW UNIVERSITY OF TECHNOLOGY


Faculty of Electronics
and Information
Technology
WARSAW UNIVERSITY OF TECHNOLOGY

Organizing Committee

- Robert Nowak, Institute of Computer Science, WUT, Warsaw (chair)
- Tomasz Gambin, Institute of Computer Science, WUT, Warsaw
- Paweł Łabaj, Małopolska Centre of Biotechnology, UJ, Kraków
- Urszula Adamiec, Institute of Computer Science, WUT, Warsaw
- Agnieszka Skalska-Bonadonna, Institute of Computer Science, WUT, Warsaw
- Katarzyna Nałęcz-Charkiewicz, Institute of Computer Science, WUT, Warsaw
- Iga Ostrowska, Institute of Computer Science, WUT, Warsaw
- Patryk Gryz, Institute of Computer Science, WUT, Warsaw
- Muhammad Farhan, Institute of Computer Science, WUT, Warsaw
- Gabriela Miączyńska, Institute of Computer Science, WUT, Warsaw

Author Index