STEP INTO THE WORLD OF SCIENCE

PTBI SYMPOSIUM 2023

SEPTEMBER 13-15, GLIWICE, POLAND

SZYB MACIEJ
Politechnika Śląska
Silesian University of Technology
WYDZIAŁ AUTOMATYKI ELEKTRONIKI I INFORMATYKI
FACULTY OF AUTOMATIC CONTROL ELECTRONICS AND COMPUTER SCIENCE
POLITECHNIKA ŚLĄSKA
FOT. SZYB MACIEJ

Book of Abstracts

Ministry of Education and Science
Doskonała Nauka II
(KONF/SP/0441/2023/01)

Politechnika Śląska

UCZELNIA BADAWCZA
INICJATYWA DOSKONAŁOŚCI

华大基因
BGI

Gliwice – przyszłość jest tu

Symposium Day 1 - 13th September, 2023

Hours	Session	Presentation title	Speaker	Affiliation	Chairperson
9:00 - 10:45	Registration/Break				
10:45 - 11:00	Welcome				
11:00 - 11:50	Keynote speaker	Long reads sequencing for the analysis of the transcriptome	Ana Conesa	Spanish National Research Council	Bartosz Wilczyński
11:50 - 12:10	Session 1	Novel clique based clustering algorithm	Sajad Shahbazi	University of Bialystok	Jerzy Tiuryn
12:10 - 12:30		Monte Carlo Feature Filtering: a Tale of a Taill	Krzysztof Mnich	University of Bialystok	
12:30 - 12:50		Developing tools to support decision-making regarding breast cancer treatment	Michał Marczyk	Silesian University of Technology	
12:50 - 13:10	Coffee break				
13:10 - 13:30	Session 2	Exploring the distribution of SNPs amongst subsequent exons and introns in the human genome	Jakub Liu	Wrocław University of Environmental and Life Sciences	Tomasz Kościółek
13:30 - 13:50		Novel expression quantitative trait loci associated with human brain glioma	Małgorzata Perycz	Polish Academy of Sciences	
13:50 - 14:10		LncRNAs Variability in Porcine Skeletal Muscle	Magdalena Frąszczak	Wrocław University of Environmental and Life Sciences	
14:10 - 14:30		Ab Initio Study of Glycine Formation In Condensed Phase	Francisco Carrascoza	Poznan University of Technology, ECBiG	
14:30 - 14:50		A novel bioinformatics tool to annotate and analyze piRNAs	Guillem Ylla	Jagiellonian University	
14:50 - 15:50	Lunch break				
15:50 - 16:30	Sponsor	Advances in De novo Assembly of Plants and Animals: T2T & Pangenome	Jianbo Jian	BGI TECH SOLUTIONS	Aleksandra Świercz
16:30 - 18:30	Poster session				
18:30 -	Further scientific discussion in restaurants/bars/pubs of Gliwice (on your own)				

Symposium Day 2 - 14th September, 2023

Hours	Session	Presentation title	Speaker	Affiliation	Chairperson
9:00 - 9:50	Keynote speaker	**The EMBO Lecture:** Knot or not? Sequence-based identification of knotted proteins with machine learning	Joanna Sułkowska	University of Warsaw	Michał J. Dąbrowski
9:50 - 10:10	Session 3	Regulatory mechanisms in the RNA World based on short RNA sequences	Jarosław Synak	Poznan University of Technology	Jacek Błażewicz
10:10 - 10:30		Application of deep generative models for RNA 3D structure prediction	Marek Justyna	Poznan University of Technology	
10:30 - 10:50		COMA - novel tool for aligning optical mapping data	Zofia Kochańska	University of Warsaw	
10:50 - 11:10	Coffee break				
11:10 - 11:30	Session 4	Microbiome health - it is time to redefine it?	Kinga Zielińska	Jagiellonian University	Małgorzata Kotulska
11:30 - 11:50		Encoding and decoding functional information of Low Complexity Regions in word embeddings vectors	Sylwia Szymańska	Silesian University of Technology	
11:50 - 12:10		Co-occurrence of amino acids in low complexity regions in proteins	Joanna Ziemska-Legięcka	Polish Academy of Sciences	
12:10 - 12:30		Sliced Wasserstein distance for creating precursor-fragment relationships in data-independent acquisition proteomics	Stanisław Grodzki	University of Warsaw	
12:30 - 12:50		Unseen passages - investigating transient tunnels facilitating ligand transport in dehalogenase	Igor Marchlewski	Adam Mickiewicz University	
12:50 - 13:00	Group photo				
13:00 - 13:50	Lunch break				
13:50 - 14:40	Keynote speaker	Unraveling the Microbial Enigma: Overcoming Bioinformatics Barriers in the Final Frontier	Kasthuri Venkateswaran	California Institute of Technology	Paweł Łabaj
14:40 - 15:00	Coffee break				
15:00 - 18:00	General Assembly of the Polish Bioinformatics Society				
19:00 - 22:00	Gala Dinner at Szyb Maciej				

Symposium Day 3 - 15th September, 2023

Hours	Session	Presentation title	Speaker	Affiliation	Chairperson
9:00 - 9:50	Honorary member	Influence of passenger mutations on expansion and extinction of cancer clones	Andrzej Polański	Silesian University of Technology	Aleksandra Gruca
9:50 - 10:10	Laureates	Development of numerical tools for screening biologically active compounds for antimicrobial effects	Mateusz Rzycki	Wroclaw University of Science and Technology	Justyna Mika & Joanna Żyła
10:10 - 10:30		GrassSV – hybrid method to detect structural variant in high throughput DNA-seq data	Domnik Witczak	Poznan University of Technology	
10:30 - 10:45		RNApdbee 3.0: webserver for 3D RNA structure analysis	Mikołaj Żurawski	Poznan University of Technology	
10:45 - 11:00		Deep neural autoencoder tool and unsupervised learning for scRNA-Seq data exploration	Anna Mrukwa	Silesian University of Technology	
11:00 - 11:10	Awards ceremony				
11:10 - 11:30	Coffee break				
11:30 - 12:20	Keynote speaker	Molecular representation learning for drug discovery	Djork-Arné Clevert	Pfizer	Anna Gambin
12:20 - 12:30	Closing remarks				
12:30 - 13:30	Lunch break				

Contents

Keynote speaker - the EMBO Lecture

Session 3

Session 4

Keynote speaker

Day 1 - 13 September 2023

Keynote speaker

Long-read sequencing for the analysis of the transcriptome

Ana Conesa[1]

[1] *Institute for Integrative Systems Biology, Spanish National Research Council*

Third-generation long-read sequencing technologies offer the potential to sequence entire transcripts and unravel the intricacies of transcriptomes. However, the analysis of long-read transcriptomics (lrRNA-seq) data presents numerous challenges. These challenges encompass distinguishing biological variability from technical noise, accurately predicting transcript models, providing precise estimates of transcript expression levels and differential expression, and elucidating the biological significance of isoform diversity. In my presentation, I will showcase the research conducted in my laboratory that addresses these challenges. Moreover, I will discuss how, through the implementation of appropriate experimental techniques and bioinformatics approaches, lrRNA-seq has the capability to unveil fresh insights into the biology of the transcriptome.

Session 1

Novel clique based clustering algorithm

Piotr Stomma[1] Witold Rudnicki[1,2] Sajad Shahbazi[2]

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

[2] *University of Bialystok, Computational Center*

We propose an efficient novel network clustering algorithm based on cliques. Method came about as a solution to a clustering task that appeared during studying relationships between microorganisms present in the human gut, using data available from the American Gut Project. Non-random patterns of co-ocurrence (or lack thereof) of different microorganisms can be detected by using mutual information (MI). We wanted to find related groups by studying cliques in network were weights are statistically significant values of MI and nodes are separate microorganisms. While there exist algorithms for maximal clique enumeration, the task of finding a partition of taxons into separate groups is not directly solved by those methods. Thus, we developed a method for building clusters using disjoint cliques as a starting point, where mean weight of connections in starting cliques is approximately maximized. Theese cliques are found by sequential greedy expansion, which exploits the correlation of the degree and strength of nodes, apparent in real data which was also found in our case. Obtained clusters were found to be biologically meaningful. Results were compared with state-of-the-art adaptive clustering methods, most important findings were confirmed by other algorithms.

We have also applied the novel approach to correlation networks obtained from gene expression data. In that case, a promising method of finding the optimal threshold of the weights in undirected graph was obtained – since original graph was too dense. Optimal threshold appears to be a maximizer of the number of nontrivial cliques found by initial clique-search.

This approach was used for more general clustering tasks, yielding a satisfactory solutions to benchmark problems, where ground truth is known. Our clustering algorithms based on initial cliques were able to solve various non trivial clustering tasks, where utilizing the threshold search consistently improved the results.

In general. our method has few hyperparameters, is applicable to various domains and appears to find clusters which have characteristics that make it useful in comparison to existing approaches, taking into the account known clustering limitations and challenges.

Monte Carlo Feature Filtering: a Tale of a Tail

Krzysztof Mnich[1] Witold Rudnicki[1,2]

[1] *Computational Centre, University of Bialystok*

[2] *Institute of Computer Science, University of Bialystok*

In bioinformatics, complex systems are investigated, that contain huge number of variables and connections. To analyse them, advanced methods are used, which involve simulations, machine learning models, network analysis, game-theory or information-theory approaches.

The analysis can lead to scores, assigned to particular variables or sets of variables. Then, the decision should be made, whether the discovery is statistically significant, given the known data. However, the null-hypothesis distribution of the score is usually not known, hence the use of Monte Carlo method is necessary. The commonly used nonparametric Monte Carlo method requires the computation of scores for a large number of artificial, random variables (called contrast or shadow variables), that follow the null distribution by design.

We evaluated the power of the Monte Carlo test in compare to a corresponding test with known null distri bution. This allowed us to estimate the number of shadow variables necessary to achieve a desired test power. The number of artificial variables turns out to be much larger than the size of the original data set, especially if a FWER correction for multiple tests is applied, which may lead to the unacceptable computational complexity.

As an alternative, we propose a parametric approach, assuming a special form of the tail of the null distribution. As well as in the nonparametric case, it is possible to evaluate the power of the parametric test depending on the number of shadow variables. We have shown, that the same power of the test can be achieved using much smaller number of artificial variables than it was necessary for the non-parametric approach.

We present two example applications of our method for analysis of biological data. The first one is a test, whether a variable takes part in any synergistic interactions with other features. The second example uses importance of a variable in a Random Forest machine learning model as a relevance metric. Here, the results may be compared with those obtained by the well-known Boruta algorithm.

Developing tools to support decision-making regarding breast cancer treatment

Michal Marczyk[1,2] Mariya Rozenblit[2] Lajos Pusztai[2] Sharon H. Giordano[3]

[1]Department of Data Mining and Engineering, Silesian University of Technology, Gliwice, Poland
[2]Department of Breast Medical Oncology, Yale School of Medicine, New Haven, CT, USA
[3]Department of Health Services Research, MD Anderson Cancer Center, The University of Texas, Houston, TX.

Breast cancer is the most frequently diagnosed cancer in the world. Typical treatment includes surgery to remove cancer tissue and chemotherapy to shrink or kill remaining cancer cells. For some subtypes, hormonal therapy could block cancer cells from getting the hormones they need to grow. Many patients and physicians struggle with decisions about adjuvant breast cancer therapy, particularly when the absolute benefits are small, and the toxicities are substantial. Thus, we developed a tool that helps to see how different treatments for invasive breast cancer can improve survival rates for individual patients after surgery. We used data from the SEER database that collects cancer incidence information together with survival data from population-based cancer registries covering around 48% of the US population. The data refer to 416,884 patients diagnosed in the years 1975-2020. Four separate Fine-Gray subdistribution hazard models with breast cancer-related death as the main event and non-cause death as the competing event were created for each cancer subtype separately: HR+/HER2-, HR+/HER2+, HR-/HER2+, and TNBC. Explanatory variables included in the final model were as follows: age, tumor size, tumor stage, and number of lymph nodes involved. Additionally, chemotherapy and radiotherapy status were added as covariates. Multivariable Fractional Polynomials (MFP) were used to model the non-linear effect of continuous explanatory variables on the response variable. Independent validation on 4,485 patients was performed using three datasets: The Cancer Genome Atlas (TCGA), Metabric, and GSE96058. The model was evaluated using two different concepts, i.e., by analyzing model calibration and its predictive ability (alive/dead) 5 years after diagnosis. The baseline model built on only 4 explanatory variables on SEER data shows very good calibration and predictive ability. By creating 4 separate models for the breast cancer subtype, we improved all performance indices compared to a single model developed using all the data. Successful validation of models on independent data proved their high performance. Finally, we developed a simple method of adding the influence of chemotherapy and/or hormone therapy on the patient's survival that allows adding any treatment regimens and showed an accurate estimation of survival. This work was partly supported by the Silesian University of Technology grant no. 02/070/BK_23/0043 for Support and Development of Research Potential.

Session 2

Exploring the distribution of SNPs amongst subsequent exons and introns in the human genome

Jakub Liu[1] Joanna Szyda[1] Magdalena Fraszczak[1] Magda Mielczarek[1]

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

Mutation in genes, including both exons and introns can have a severe impact on downstream products of gene expression. A common measure of genetic modifications between a sequence and its reference are single nucleotide polymorphisms (SNPs). This study aimed to answer the question if SNPs are evenly or unevenly distributed in subsequent exons or introns. The genomic sequence of the first and 22nd human chromosomes were investigated, since the first chromosome, is the largest human chromosome and the 22nd chromosome represents smaller human chromosomes. The information about the polymorphisms was stored in the standard Variant Call Format (VCF). The raw VCF files were filtered using the VCFtools software to only obtain variants of desired high quality The genomic annotation of the SNPs was obtained using the Ensembl Variant Effect Predictor tool. The next step was to filter out only variants that are in introns or exons and that are found in canonical transcripts. Subsequently, genes were classified into categories corresponding to their total number of exons or introns. For every gene and separately for introns and exons, the number of SNPs that fall into the above- mentioned categories was counted. In the next step, all genes that have the same number of introns or exons were grouped together. This allowed to compare the number of variants in subsequent introns/exons. For example, for all the genes that have 3 introns, we could see how many SNPs are in introns 1/3, 2/3 and 3/3. Subsequently, the Friedman test was applied to check whether the SNP distribution in every category is uniform. Within this test, the gene ID was set as the blocking variable to account for genes with extremely high SNP density. For genes in which the null hypothesis of a uniform SNP distribution was rejected, a post-hoc test was utilized to see between which exons/introns significant differences in SNP counts occur. The results show that there is a significantly higher number of SNPs in the first intron. This is true for genes that have from two to nine introns. Contrariwise, there is significantly more SNPs residing in the last exon. This trend holds true for genes that have up to 5 exons.

Novel expression quantitative trait loci associated with human brain glioma

Małgorzata Perycz[1,5] Ilona Grabowicz[1,5] Paulina Szadkowska[2] Bartosz Wojtas[2] Bartek Wilczyński[3] Tomasz Żok[4] Marta Jardanowska[1] Bożena Kamińska[2] Michał J. Dabrowski[1]

[1]*Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland*
[2]*Nencki Institute of Experimental Biology, Polish Academy of Sciences, Pasteura 3, 02-093 Warsaw, Poland*
[3]*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Stefana Banacha 2, 02-097, Warsaw, Poland.*
[4]*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland*
[5]*Contributed equally*

Gliomas are primary malignant brain tumors of high heterogeneity and invasive profile, which contribute to their poor response to available therapies and limit treatment options. We used a multi-step analysis pipeline with stringent filtering to identify novel expression quantitative trait loci associated with differential gene expression between glioma grades. Based on the RNA-seq data obtained from 108 resected tumors (11 pediatric astrocytoma, 31 diffuse astrocytoma and 66 glioblastoma and 1 pediatric glioblastoma) from the Polish patients, we identified a group of actively transcribed genes, in which we found variants, whose penetration depth highly correlated with gene expression levels. The variants present in genes differentially expressed between glioma grades were selected. We next used DNA amplicon sequencing of 5 of SNVs present in VEGFA and DDX11 gene products. In total, we have identified 83 mutations (substitutions / indels) in 52 genes, whose penetration depth strongly correlated with the levels of gene expression. We conclude that these eQTLs may account for some of the heterogeneity observed between glioma grades and thus should be considered as a target for further research.

LncRNAs Variability in Porcine Skeletal Muscle

Bartłomiej Hofman[1] Magdalena Fraszczak[1] Joanna Szyda[1,2] Magda Mielczarek[1,2]

[1]Biostatistics group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw 51-631, Poland
[2]National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

Long non-coding RNAs (lncRNAs) represent a class of transcribed RNA molecules that are longer than 200 nucleotides and do not encode proteins. lncRNAs play a role in gene expression regulation by affecting transcription, DNA replication, and the response to DNA damage and repair. In the current study, we analised the interindividual variability of lncRNA in swine as well as identified the possible lncRNAs target genes. Transcriptomes were collected from muscle tissue (longissimus dorsi) samples, which were taken from six Polish Landrace boars, where three boars were half-brothers. RNA fragments were sequenced using the Illumina NovaSeq 6000 in a paired-end mode with the designated single read length of 101 bp. The number of raw reads per individual ranged from 241,605,646 to 337,017,230. The total number of common lncRNAs among all boars was 232. The number and length of lncRNA differed significantly between individuals. Moreover, significant inter-individual variability in the expression level was observed. There is no linear correlation between the lncRNAs length and their expression level. Therefore, it can be excluded that the expression was biased by some extraordinary long/short lncRNAs. Both, co-localized and co-expressed target genes enrichment analysis determined a variety of biological processes which play fundamental role in the cell biology. There were mostly related to response to stress, metabolic processes, translation, ribosome biogenesis, protein regulation and the dynamics of multiple DNA- and RNA-protein complexes.

Ab Initio Study of Glycine Formation In Condensed Phase

Francisco Carrascoza[1,2] Piotr Lukasiak[1,2] Wieslaw Nowak[3] Jacek Blazewicz[1,4]

[1]Institute of Computer Science, Poznan University of Technology, Poznan, Poland

[2]European Centre for Bioinformatics and Genetics ECBiG, Poznan, Poland

[3]Institute of Physics, Faculty of Physics, Astronomy and Informatics, N. Copernicus University, Torun, Poland

[4]Institute of Bioorganic Chemistry, Poznan Academy of Sciences, Poznan, Poland

Glycine's role in protein and prebiotic substance formation is widely acknowledged. However, understanding the mechanism behind its spontaneous formation under prebiotic Earth conditions and within the interstellar medium (ISM) is still a subject of debate, given the Earth's changing geochemical environment over time. Detecting glycine within the ISM presents challenges as well, although its formation in water-rich ice grains within dense molecular clouds is believed to be possible. In this study, we employed ab initio molecular dynamics (MD) simulations enhanced with modern free energy calculations to model the chemical reaction involving carbon monoxide, formaldimine, and water to produce glycine. We investigated the conditions under which glycine forms in the condensed phase at temperatures of 50K, 70K, 100K, and 300K. Additionally, we examined the impact of different electric fields on this process. Our findings demonstrate that glycine can be formed with remarkably low energy barriers of 0.5 kcal/mol at 50K. This study aims to explore whether this reaction could serve as a viable mechanism for glycine formation in the ISM and on planet Earth. By doing so, we contribute to the ongoing search for a consensus among various proposals regarding glycine formation. Moreover, our work underscores the significance of metadynamics and Car-Parrinello MD methods as valuable tools in unraveling complex, multistep reaction pathways that may play a crucial role in astronomical phenomena.

A novel bioinformatics tool to annotate and analyze piRNAs

Dominik Robak[1] Guillem Ylla[1]

[1]Laboratory of Bioinformatics and Genome Biology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University

Piwi-interacting RNAs (piRNAs) are small RNAs that play significant roles in safeguarding the genome against transposable elements (TE) in animal germ cells. Recent studies have also revealed instances of piRNAs regulating the expression of endogenous genes, thereby impacting various biological functions. Interestingly, we have also observed piRNAs highly expressed in somatic tissues of different basal animal lineages such as insects. These observations have raised many questions about the evolution and diverse roles of piRNAs in animals.

However, studying piRNAs poses significant challenges, particularly in their accurate identification, annotation, and classification. To address these issues, we developed a Python package that leverages small RNA-seq data mapped to a genome to identify and classify piRNAs produced either by the primary or secondary pathways. In the primary piRNA biogenesis pathways, long precursor RNAs are transcribed and subsequently cleaved into the small piRNA (26-28nts) that are bound to PIWI family proteins. The secondary biogenesis pathway (aka. ping-pong pathway) uses the RNA molecule of the primary piRNA targets to produce secondary piRNAs which will bind to the piRNA precursor and amplify the production of the primary piRNA.

Our tool identifies both, genomic piRNA clusters which produce the primary piRNAs as well secondary piRNA loci. The piRNA primary clusters are identified based on an adapted version of the density-based clustering algorithms for data with noise DBASCAN. The secondary piRNA loci are identified based on the ping-pong signature, characterized by pairs of small RNA-seq reads mapped in complementary strands of the same genomic locus with 10 nucleotide overlap at their 5'ends. In addition to annotating the piRNAs, our tool produces informative graphical reports, and other files readily formatted for common downstream analysis. The tool has been shown to outperform current tools in the different animal species tested.

In summary, we present a powerful bioinformatics tool to annotate, classify, and analyze piRNAs in animals using small RNA-seq data aligned to a reference genome. This tool not only addresses the limitations of current approaches but also allows us to provide valuable insights into the diverse roles and evolutionary significance of piRNAs in animal biology.

Sponsor

Advances in De novo Assembly of Plants and Animals: T2T & Pangenome

Jianbo Jian[1]

[1] *BGI Tech Solutions, Warsaw, Poland*

In the past few decades, de novo assembly has witnessed rapid development. Numerous reference genomes of plants and animals have been successfully completed. Recently, Telomere-to-telomere (T2T) genomes have emerged as high-quality complete genomes with exceptional genomic accuracy, continuity, and integrity. They play a crucial role in unraveling variations in centromeres and evolutionary patterns of complex structures such as telomeric regions. However, the limitations of single species reference genomes in detecting genetic variations have led to the emergence of the concept of pan-genomes. In this presentation, we will provide a comprehensive overview of recent advancements in plant and animal genomics. Specifically, we will showcase several noteworthy cases published by BGI, including Paeonia osti (tree peony), Chinese tapertail anchovy, and East Asian finless porpoise. Furthermore, we will present the High-resolution silkworm pan-genome and sorghum pan-genome projects.

BGI Genomics is one of the world's leading genomics organizations. It provides sequencing, mass spectrometry based proteomics and metabolomics, bioinformatics services, and bio-cloud computing.

Posters

The KFERQminer Software – development of novel bioinformatics tools for finding chaperone-mediated autophagy motifs

Marcin Lubocki[1] Anna Szczepińska[1] Dzmitry Dauhalevich[1] Maria Borysiak[1] Andrea Lipińska[1]

[1] *Laboratory of Virus Molecular Biology, Intercollegiate Faculty of Biotechnology, University of Gdańsk, Medical University of Gdańsk*

Chaperone-mediated autophagy (CMA) is a conserved mammalian process that selects proteins with KFERQ-related motifs in their sequence for lysosomal degradation. Despite knowing about this process for decades, no dedicated bioinformatics resource is still integrating knowledge about degraded proteins and KFERQ motifs. To bridge this gap, we have developed the KFERQminer Database, which allows the exploration of already described in literature KFERQ motifs. The collection is continuously updated and involves a thorough manual review of papers published in PubMed, searching for CMA-related keywords to extract and summarize the existing data. The collected set is subsequently used for sequence analysis and the development of new in silico KFERQ motif searching tools. The integrated tool enables searching for new motifs in protein sequences and has several advanced features that are not present in existing software. These include (1) the use of mathematical scoring to predict motifs, (2) the comparison of newly discovered motifs with those described in the scientific literature, and (3) the integration of external deep learning tools for protein analysis, which strengthens predictions by embedding them alongside relevant biological information. In the final step, proteins with the most promising motif targets are validated as potential CMA substrates in cell line models. Moreover, as intracellular pathogens interact with the cellular processes of their hosts, we have analyzed proteomes of human viruses using the KFERQminer software. Interestingly, we observed an abundance of potential KFERQ motifs in several virus families, which may have significant medical implications. Supported by several bioinformatics analyses, our observations suggest the legitimacy of extending the search for KFERQ motifs to intracellular pathogens of CMA-competent organisms, which, by their nature, are subjected to the cellular processes of their hosts.

Novel clique based clustering algorithm

Piotr Stomma[1] Witold Rudnicki[1,2] Sajad Shahbazi[2]

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

[2] *University of Bialystok, Computational Center*

We propose an efficient novel network clustering algorithm based on cliques. Method came about as a solution to a clustering task that appeared during studying relationships between microorganisms present in the human gut, using data available from the American Gut Project. Non-random patterns of co-ocurrence (or lack thereof) of different microorganisms can be detected by using mutual information (MI). We wanted to find related groups by studying cliques in network were weights are statistically significant values of MI and nodes are separate microorganisms. While there exist algorithms for maximal clique enumeration, the task of finding a partition of taxons into separate groups is not directly solved by those methods. Thus, we developed a method for building clusters using disjoint cliques as a starting point, where mean weight of connections in starting cliques is approximately maximized. Theese cliques are found by sequential greedy expansion, which exploits the correlation of the degree and strength of nodes, apparent in real data which was also found in our case. Obtained clusters were found to be biologically meaningful. Results were compared with state-of-the-art adaptive clustering methods, most important findings were confirmed by other algorithms.

We have also applied the novel approach to correlation networks obtained from gene expression data. In that case, a promising method of finding the optimal threshold of the weights in undirected graph was obtained – since original graph was too dense. Optimal threshold appears to be a maximizer of the number of nontrivial cliques found by initial clique-search.

This approach was used for more general clustering tasks, yielding a satisfactory solutions to benchmark problems, where ground truth is known. Our clustering algorithms based on initial cliques were able to solve various non trivial clustering tasks, where utilizing the threshold search consistently improved the results.

In general. our method has few hyperparameters, is applicable to various domains and appears to find clusters which have characteristics that make it useful in comparison to existing approaches, taking into the account known clustering limitations and challenges.

How data-driven methods advance Decision Tree optimisation for heterogeneous data: a comparison

Viviana Laber[1] Joanna Polanska[1]

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

When handling non-homogeneous data, the choice of the appropriate Decision Tree optimisation method poses a challenge to be overcome. Our study addresses this problem by developing a new approach to optimising Decision Trees which considers the heterogeneity of the data, and compares it with existing optimisation methods: cost-complexity pruning and feature selection using the genetic algorithm.

We employed a scoring system that reduces data heterogeneity through significantly limiting the feature space and ran it on a dataset consisting of 4,795 samples with 261 radiomic features each. The samples came from 2,424 healthy donors, 1,136 non-COVID pneumonia patients and 1,235 COVID-19 pneumonia patients. This method utilised the decision tree structure to provide a better understanding of the data. Based on the samples' decision paths, every instance of a node visit was counted and summed up according to the feature it was associated with, resulting in an importance score for each feature. New sets of features that presented importance significantly larger than the majority were extracted using Youden's index. A global set of 15 features and a set of similar size specific to each class were obtained.

To compare our new approach with the traditional methods, we conducted k-means cross-validation and additionally evaluated their performance with a separate balanced holdout set which was the result of a random split of the original dataset. Both complete and optimised tree structures were generated for further analysis.

As indicated by a PPV of 52.09% (CI 42.41%-50.04%) and NPV of 76.40% (71.38%-77.21%) for the global feature subset and a PPV of 56.99% (43.77%-49.83%) and NPV of 78.65% (71.97%-77.16%) for the COVID feature subset, our approach has achieved results comparable to cost-complexity pruning with a PPV of 53.51% (39.52%-58.01%) and NPV of 76.82% (71.53%-81.06%). Additionally, as the PPV and NPV values of our system remain at a satisfactory level, it has a high potential for use in screening tests. The feature set obtained through the genetic algorithm exhibited a relatively poorer performance with a PPV of 45.26% (40.77%-49.77%) and an NPV of 72.33% (71.12%-76.51%). Although there is little overlap between the genetic algorithm feature set and our method's sets of features, many of the features exhibit moderate to high correlations.

COMA - novel tool for aligning optical mapping data

Norbert Dojer[1] Mikołaj Arciszewski[1] Zofia Kochańska[1]

[1] *University of Warsaw*

Advancements in DNA sequencing technologies have revolutionized the field of genomics, enabling researchers to generate vast amounts of genomic data at an unprecedented scale. Among these technologies, optical mapping has emerged as a promising approach for long-range genome analysis. It leverages the power of fluorescence microscopy to directly visualize and map DNA molecules, providing valuable insights into structural variations and genome organization.

Accurate alignment of optical mapping sequences is crucial for interpreting and extracting meaningful biological information. However, due to the unique characteristics of optical maps traditional sequence alignment approaches often fall short in providing accurate alignment results. To address this challenge, we present a novel tool called COMA (Cross-correlation Optical Map Alignment), specifically developed for aligning optical mapping sequences and overcoming these difficulties.

COMA tackles this problem by incorporating a double cross-correlation approach that leverages the specific features of optical mapping data. It utilizes two separate computations of cross-correlation to first identify potential locations where a molecule could be mapped and then perfects the mapping. The tool also incorporates an extensive set of parameters that can be adjusted to fit the currently studied data and modify its stringency. What makes it even more unique is the additional output of locations where there were noticeable conflicts in alignment. These discrepancies can provide additional information on potential genomic regions where structural variants can be observed.

Additionally, we investigate the impact of COMA on downstream analysis tasks, such as structural variant detection. We compare our results with those obtained by existing tools. We have created a novel tool which identifies structural changes and verifies them against the existing benchmark dataset, which serves as a gold standard. It is adapted to use a file format that is standard for aligned optical maps, allowing it to be used on results obtained using different tools.

In conclusion, this study presents COMA, a novel tool specifically designed for aligning optical mapping sequences. We believe that COMA will serve as a valuable resource for researchers working in the field of genomics, offering new opportunities to gain deeper insight into the structure of genomes.

Effect size profiles for analysis of multi-species single cell RNA-seq data

Anna Papiez[1] Jonathan Pioch[2] Hans-Joachim Mollenkopf[3] Björn Corleis[2] Anca Dorhoi[2] Joanna Polanska[1]

[1]Department of Data Science and Engineering, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[2]Institute of Immunology, Friedrich Loeffler Institute, Südufer 10, 17493 Greifswald - Insel Riems, Germany
[3]Department of Immunology, Max Planck Institute for Infection Biology, Virchowweg 12, 10117 Berlin, Germany

The integration of data from experiments conducted on multispecies studies makes invaluable contributions to the comprehension of fundamental disease mechanisms that traverse species boundaries. However, analyzing gene expression across multiple species is often challenging due to inconsistencies in annotations and when dealing with limited sample sizes, which can introduce biases stemming from batch effects. In this study, we present an alternate approach to the standard statistical inference in single cell RNA-sequencing that is well-suited for merging data from small samples and multiple organisms.

The analysis pipeline relies on effect size metric profiles of samples within specific cell clusters. Instead of inferring from conventional differentiation analyses based on p-values, we utilize effect size metrics as a substitute. The profiles generated using these effect size metrics serve as a valuable tool for establishing connections between cell type clusters across the organisms under study. Our algorithms were validated using data obtained from human and bovine peripheral blood mononuclear cells stimulated with Mycobacterium tuberculosis. The challenges included the availability of only single samples per condition, which would render classic p-value based statistical analysis unfeasible. We employed effect size ratios to identify genes that exhibited differential expression in control and stimulated samples. The genes identified through effect size profiling were further confirmed experimentally using qPCR.

Our findings demonstrate that in situations where batch effects exert significant influence on cell type variation in multispecies studies with limited sample sizes, effect size profiling emerges as a valid alternative to traditional statistical inference techniques.

This work is supported by the Silesian University of Technology grant no. 02/070/BK_23/0043 for Support and Development of Research Potential.

Revealing the RBP regulome in hepatocellular carcinoma via consensus GRN inference

Mateusz Garbulowski[1] Riccardo Mosca[2] Carlos J. Gallardo-Dodd[2] Claudia Kutter[2] Erik L. L. Sonnhammer[1]

[1]Department of Biochemistry and Biophysics, Stockholm University, Science for Life Laboratory, Solna, Sweden
[2]Department of Microbiology, Tumor, and Cell Biology, Karolinska Institute, Science for Life Laboratory, Solna, Sweden

RNA binding proteins (RBPs) play a key role as regulators of diverse gene expression mechanisms by binding RNA targets. RBPs are associated to post-transcriptional functions in cells such as mRNA decay, splicing, stabilization, translation and transport. In cancer, they are involved in mechanisms of drug resistance and oncogenesis. Therefore, it is vital to explore the RBP regulome in cancer. In this study, we aim at determining the hepatocellular carcinoma (HCC) regulome. In order to characterize such relationships, we employed the gene regulatory network (GRN) inference approach. In this work, the GRN inference relies on the wisdom of crowds concept that assumes creating a consensus GRN using multiple inference methods. Fundamentally, to create the GRN we used experiments that allow for controlled perturbation of gene expression, i.e., shRNA knockdown followed by RNA-seq (shRNA-seq). Thus, the consensus GRN is built upon 12 regression and machine learning methods in so-called perturbation-based fashion. To evaluate the quality of our methodology, we benchmarked the consensus approach on a synthetically created data. Afterwards, we investigated the ENCODE HepG2 shRNA-seq data. To investigate HCC consensus GRN, we performed a comprehensive validation including such resources as GTEx and TCGA RNA-seq, ENCODE eCLIP-seq, FunCoup5 scores, in-house RAP-seq and overlap with Gene Regulatory Network Database (GRAND). Finally, on the validated GRNs we performed community detection and gene enrichment analysis. In conclusion, the benchmarking uncovered that a consensus approach reduces the number of false positive links in contrast to single GRNs. Moreover, we identified a list of 130 RBP-RBP relationships related to HCC. Among others, we found that LIN28B positively regulates PTBP1, what was confirmed with GRAND, eCLIP-seq and a high support of consensus links. Furthermore, the expression level of LIN28B significantly impacts a survival of HCC patients. Lastly, the community enrichment revealed subnetworks related to various cancer and survival mechanisms.

Reliability of Oxford Nanopore flowcell reuse on taxonomic identification of environmental microbial communities with short-read sequencing validation in whole metagenomic sequencing

Dedan Githae[1] Agata Jarosz[1] Kinga Herda[1] Kamila Marszałek[1] Wojciech Branicki[2] Paweł Łabaj[1]

[1]Małopolska Center of Biology, Jagiellonian University, Krakow, Poland

[2]Institute of Zoology and Biomedical Research, Jagiellonian University, Krakow, Poland

Background: Emergence of high throughput sequencing technologies has greatly enabled researchers to study microbiomes at much deeper levels than before. They are broadly classified as second generation (short-read sequencing) and third generation (long-read sequencing). While second generation sequencing of short-read sequencing gained popularity earlier and is widely used, it bears setback associated with sequence assembly of these short reads into longer contigs in absence of reference sequences for mapping. The third generation sequencing of long reads bypass this bottleneck since they are capable of sequencing entire small genomes spanning several KBs.

In this study, we aimed at identification of key abundant microbial taxa in different ecological niche, using both long read (Oxford Nanopore Technology [ONT]) and short read (Illumina) sequencing platforms. We also aimed at assessing the capability of ONT flowcells in identifying sequencing key taxa, and how reliable they are despite washing and repeated re use for subsequent runs due to membrane deterioration and reduced pore activity. We also aimed at evaluating the consistency of the identified key taxa between these two platforms and implications in forensic use.

Methodology: Four samples, previously sampled from different environments were barcoded and sequenced using four ONT flowcells. The sequencing order of sample being sequenced was switched among the flowcells upon washing after each run to monitor overall throughput per sample at different washing steps across the flowcells. The long reads produced per sample in each flowcell were then basecalled, and extracted by their respective barcodes for taxonomic identification. For short read taxonomic analysis, k-mer based approach was used to assign taxonomy.

Conclusion: From this study, we were able to identify that despite reduction of sequencing capability across different sequencing runs due to increased unavailable and inactive pores, taxonomic identification of the taxa remained consistent in the samples, with loss of only lowly abundant genera. We also identified that flowcell washing still had sequencing residue from previous runs, a factor which was overcome by using unique barcodes.

How to encode in no-mapping mapping of Oxford Nanopore long reads?

Tomasz Strzoda[1] Lourdes Cruz-Garcia[2] Mustafa Najim[2] Christophe Badie[2] Joanna Polanska[1]

[1]Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland,
[2]Cancer Mechanisms and Biomarkers Group, Radiation Effects Department, Radiation, Chemical Environmental Hazards, UK Health Security Agency, Didcot, UK

Sequence mapping is an important and time-consuming step in the processing of RNA/DNA sequence data. In the study carried out, we replaced this process with a machine learning model using methods known from natural language processing (NLP) and analysed the different ways of encoding sequences. Six RNA samples were available, representing in total 8.5 million Oxford Nanopore long reads. First, the entire strand was represented as words of 3 or 6bp with one nucleotide offset. Then 1- or 3-word combinations were created. Finally, the following encodings were studied: '3bp', '3bp—3bp—3bp', '6bp—6bp—6bp', for which the bag-of-words method was used. The architecture of the classifier was based on a 3-layer neural network of 50, 50 and 1 neurons, respectively. To evaluate the model Leave-A-Sample-Out-Cross-Validation scheme was used, allowing each of the six samples to be taken as a test set. All of the proposed approaches achieved high scores on the classifier evaluation metrics. The worst of these, '3bp', obtained 94.50% accuracy, which was to be expected, due to the very short word and the disregard of other neighbouring words. The best, '6bp—6bp—6bp', reached 98.29% accuracy, but was only 0.77% better than '3bp—3bp—3bp' and contained a 64x larger feature domain than it. These results confirmed the potential of using NLP methods in bioinformatics and showed the impact of encoding approaches on the final classifier performance.

Investigation of microbial intra-community synergies as an improvement of soil biome understanding

Dagmara Błaszczyk[1,2] Witold Wydmański[1,2] Krzysztof Mnich[3] Kinga Zielińska[1] Valentyn Bezshapkin[1,4] Michał Kowalski[1,2] Alina Frolova[1] Renata Zbieć-Piekarska[5] Wojciech Branicki[1] 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] *ETH Zurich, Institute of Microbiology, Zurich, Vladimir-Prelog-Weg 4*

[5] *Central Forensic Laboratory of the Police, Warszawa, Al. Ujazdowskie 7*

[6] *University of Białystok, Institute of Computer Science, Bialystok, Konstantego Ciołkowskiego 1M*

Microbiota research is increasingly focused on the exposome factors and their relation to metadata. It allows for finding microorganisms that are specific to ecological niches. However, most study examines microorganisms found in a sample as separate features, which does not give the whole picture of interactions between microorganisms in environmental niches. Our project aims to discover the synergies between microorganisms and examine their impact on the classification of samples in one of the established Polish microclimate clusters. In our study, we use 240 soil samples collected from different locations in Poland which have been sequenced with extreme depth of over 100M paired-end reads (Whole Metagenome Sequencing). The sampling locations have been selected based on climate characteristics supported by over 20 years of weather conditions parameters history and represent three different Polish microclimate clusters. We use MDFS (MultiDimensional Feature Selection) (Mnich Rudnicki, 2020), based on the 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 first results indeed confirm the existence of microclimate-related and local-specific microbial communities, which is in line with earlier studies of the MetaSUB Consortium (Danko et al., 2021) on a global scale. Moreover, by examining synergies we are able to reduce the number of features while maintaining a similar level of classification. Just with Poland being very homogeneous from a climate and biome perspective, we investigate whether microbial synergies might be the key to studying the diversity of microorganisms between closely related ecological niches.

Gut microbiome is associated with fatal outcome and ICU admission in patients with COVID-19

Katarzyna Kopera[1] Tomasz Gromowski[1] Witold Wydmański 1[5] Karolina Skonieczna-Żydecka[2] Agata Muszyńska[1] Kinga Zielińska[1] Anna Wierzbicka-Woś[3] Mariusz Kaczmarczyk[2,3] Roland Kadaj-Lipka[4] Wojciech Marlicz[2,3] Igor Łoniewski[3] Paweł P. Łabaj[1] Grażyna Rydzewska[4] Tomasz Kosciolek[1]

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

[2] *Pomeranian Medical University, Szczecin, Poland*

[3] *Sanprobi Sp. z o.o. Sp. k, Poland*

[4] *Central Clinical Hospital of the Ministry of Interior and Administration, Warsaw, Poland*

[5] *Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland*

Gastrointestinal symptoms have been reported in addition to the commonly observed respiratory symptoms of COVID-19. The gut microbiome plays a crucial role in various immunological and metabolic pathways that affect both illness etiology and overall health. This effect of the microbiome on the course of the disease and health management was demonstrated in COVID-19. Our study involved shallow and deep shotgun sequencing methods on 347 hospitalized COVID-19 patients and healthy controls to track changes in microbiota composition induced by COVID-19 treatment and correlate these alterations with clinical outcomes such as ICU admission and mortality. Using machine learning we sought to find if the microbiome is predictive of COVID-19 prognosis and if prediction based on microbiome outperforms baseline classifiers. Finally, we evaluated the taxonomic agreement between shallow shotgun and deep shotgun sequencing and validated shallow shotgun sequencing application in COVID-19 clinical setting.

Analysis of the CNV detection based on combined Illumina and Nanopore data

Paulina Leśniak[1,2,3] Magda Mielczarek[1,2,3,4] Błażej Nowak[1,2,3] Joanna Szyda[1,2,3,4] Tomasz Strzała[1,2,3] Magdalena Fraszczak[1,2,3]

[1] *Biostatistics Group,*

[2] *Department of Genetics,*

[3] *Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland*

[4] *Institute of Animal Breeding, Krakowska 1, 32-083 Balice, Poland,*

In this study, we compared various copy number variation (CNV) detection methods using, short Illumina and long Oxford Nanopore reads. CNV discovery based on short reads is challenging due to the complex genome structure, especially in repeated regions. The increasing availability of long-read technologies like Oxford Nanopore Technology has greatly facilitated CNVs discovery, however, these technologies remain too costly to apply routinely to population-level studies and it is characterised by the high error rate. Integrating short and long reads is beneficial as they can mutually mitigate their limitations. Data used for this project consisted of the whole genomic DNA of 6 swines representing the Polish Large White breed obtained using the Illumina HiSeq2000 platform and Oxford Nanopore Technology. Four CNV detection tools were used: CNVnator and Pindel in the case of Illumina data and Sniffles and CuteSV for long read sequences. The total number of identified variants varied between 27 and 111 in the case of CuteSV and 211720 and 309924 for Pindel. In order to conduct the comparison we performed diverse statistical tests, especially 2, Friedmann and signed rank Wilcoxon tests We observed that the choice of technology influenced the obtained results. Deletions and duplications detected using CNVnator are significantly longer than those identified by other tools, although the longest deletion (duplication) equal to 11608700 (47807962) was identified using Sniffles. Only between 0.05% (0.06%) and 2.27% (0.54%) of all founded deletions (duplication) were common for both methods. As the last step distribution and density of CNVs along the genome were investigated, based on polymorphism identified in both methods. Variants were mostly located in intergenic regions and their location is non-randomly distributed.

AlphaFold reveals kingdom specific relationships between low complexity regions and structural disorder in proteins

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

[1] *Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland*

[2] *Department of Computer Networks and Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology Gliwice, Poland*

Low Complexity Regions (LCRs) are fragments of protein sequences characterized by low amino acid diversity. In many studies LCRs have been linked to disordered protein structure, however comprehensively investigating the relationship between LCRs and protein structure has always been challenging due to the biased nature of protein structure databases, which primarily included globular proteins. Recent development of the AlphaFold method have allowed for providing a broader perspective on this issue, as the predicted structures now cover about 80% of the proteome. To better understand the role of LCRs in protein 3D structure formation and therefore their function, we decided to analyze the distribution of the values of AlphaFold predicted Local Difference Distance Test (pLDDT) measure among identified LCRs for selected kingdoms of organisms. Additionally, we investigated the relationship between pLDDT values for human LCRs and intrinsically disordered regions (IDRs) collected in the DisProt database. Since the pLDDT measure represents the confidence score of the AlphaFold method in predicting the 3D structure, it can also be used as a predictor of structural disorder. To determine a specific region as LCR, we used three approaches: the SEG algorithm with two sets of parameters and the CAST algorithm. The results show that considering only the kingdom level of taxonomy there is a predominance of regions with low pLDDT value (<50) among Fungi and Plants. Bacterial LCRs demonstrated higher pLDDT values. Results for subphylums Invertebrates and Vertebrates showed many similarities regardless of the LCR detection method and had higher values of the prediction measure. Human LCRs are characterized by lower pLDDT measure than Vertebrates. We also showed that for human disordered regions the average pLDDT value is higher than human for low complexity regions. We analyzed the distribution of pLDDT values, understood as a measure of structural disorder in LCRs. In some groups of organisms the relationship between low pLDDT values of sequence fragments and its low complexity is vividly noticeable. Bacterial LCRs are rich in proline and glycine which allows them to form flexible but stable secondary structures such as helices. AlphaFold algorithm is able to more accurately predict such structures and this may be reflected in very high pLDDT values for such organisms as it is a prediction confidence measure.

Using single-cell transcriptomics to compare functional properties of repopulated and naïve microglia after pharmacological depletion

Zuzanna M. Luczak-Sobotkowska[1] Patrycja Rosa[1,2] Maria Banqueri[1] Natalia Ochocka[1] Aleksander Jankowski[2] Bozena Kaminska[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*

Microglia are heterogenous population of myeloid cells residing in the central nervous system. They are responsible for maintaining homeostasis in brain and preventing potential damage to other cells. Previous studies defined transcriptional profiles of individual microglia subpopulations and provided insights in their function. Interestingly, microglia are one of the populations of brain cells which can fully renew during the lifetime of an organism. In this study, we investigated the origin and functionality of repopulated microglia in mice, using single-cell RNA sequencing (scRNA-seq).

Pharmacological treatment with BLZ-945 resulted in almost complete microglia depletion (¡1

Comparisons between main microglial clusters allowed us to point out the most interesting genes associated in the repopulation processes. These genes were involved in regulatory pathways and ontologies responsible for cell differentiation and maturation. In addition, we compared microglia repopulation in young (3 months) and older (12 months) mice, and observed consistent cell clustering. Surprisingly, we noticed that older repopulated microglia lacked mature microglial cells, suggesting that most cells were quickly programmed to apoptosis and induced cell death.

Deep Functional Residue Information (DeepFRI) enabled to identify adaptation to space conditions in International Space Station microorganisms.

Lukasz Szydlowski[1] Anna Simpson[2] Nitin Singh[2] Deniz Ece Kaya[3] Alper Bulbul[4] Osman Ugur Sezerman[4] Tomasz Kosciolek[1] Paweł P. Łabaj[1] Kasthuri Venkateswaran[2]

[1] Malopolska Center of Biotechnology, Jagiellonian University, Krakow, Poland

[2] California Institute of Technology, Jet Propulsion Laboratory, Pasadena, United States

[3] King's College London, London, United Kingdom

[4] Acibadem University, Istanbul, Turkey

The harsh conditions of outer space create unique selective pressures on microorganisms. This study focuses on the functional annotation of seven Gram-positive bacterial isolates derived from the International Space Station (ISS) and Jet Propulsion Laboratory-Spacecraft Assembly Facilities (JPL-SAF) during the Mars 2020 mission, including a representative of a new genus. Using genome assembly and the machine learning-based functional annotation tool Deep Functional Residue Information (DeepFRI), as well as sequence-based homology and orthology analyses, we compared the predicted functional characteristics of these microorganisms with their closest Earth-bound relatives, including genes associated with radiation resistance, microgravity adaptation, stress response, and metabolic rearrangements. By analyzing the genomes and possible protein-coding sequences of these isolates, we have identified common features associated with adaptations to space conditions. These adaptations include the use of mechanosensitive channel proteins to mitigate microgravity-related hypoosmotic stress, DNA repair systems and mobile genetic elements to combat increased radiation exposure. Our study demonstrates the superior coverage of DeepFRI's functional annotations compared to homology-based tools and highlights the potential of knowledge-based genome mining to enhance our understanding of previously-uncharacterized microbial adaptation to extreme conditions Our findings also provide a set of biomarkers that could aid in astrobiological studies targeting life on other planets.

Effect size for unsupervised pathway enrichment analysis in scRNA-Sequencing

Kamila Szumala[1] Joanna Zyla[1]

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

Pathway enrichment (PE) is one of the leading steps in bioinformatical analysis of high-throughput molecular biology data. With the development of scRNA-Seq, the single-sample PE algorithms became more popular as they can grab the pathway activity of each cell. Another challenge of scRNA-Seq data is cell labelling. In the presented work, the ability of significant pathways detection using single-sample PE was checked using two different approaches: statistical inference and its combination with unsupervised learning. The PBMC dataset of 3222 cells and the expression levels of 15817 genes was used. The cells were initially labelled into six leading types. The log normalization in the Seurat package was applied. Next, gene expressions were transformed into pathway activity scores (PAS) using explainable sets of genes representing cell type signatures of PBMC. Gene sets were extracted from publicly available databases i.e. Cell Marker, CYBERSORT, PanglaoDB, KEGG and tmod (in total 204 pathways). Transformation into PAS was performed by the six widely known single-sample PE methods (CERNO, AUCell, PLAGE, GSVA, Z-score and Mean). For each obtained PAS, the Louvain clustering was applied to determine unknown groups of cells. Next, two-sample t-test was performed between obtained clusters (cluster vs rest) to seek the significant pathway. Obtained p-values were corrected by the Bonferroni method. A Cohen d effect size (ES) was calculated in the same manner, and Gaussian mixture decomposition (GMM) was applied to extract the most active pathways for each cluster. Using statistical inference, with strict Bonferroni correction, over 85Standard statistical inference make interpretation very difficult and almost impossible to identify the cell type in cluster. The Cohen d combined with GMM can effectively indicate significant pathways within predefined clusters. Moreover, obtained results define the cell types in particular clusters for most investigated algorithms. Financed SUT grant for maintaining and developing research potential.

Three longitudinal regimes of the human gut microbiome

Zuzanna Karwowska[1] Paweł Szczerbiak[1] Tomasz Kosciolek[1,2]

[1]Bioinformatics Group, Malopolska Center of Biotechnology Jagiellonian University
[2]Department of Data Science and Engineering, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland

Despite the majority of microbiome studies being cross-sectional, it is widely acknowledged that the microbiome is a dynamic ecosystem. Here, we analyse how the gut microbiome changes over time as a community, how different bacterial species behave over time, and whether there are clusters of bacteria that exhibit similar fluctuations. We show that a healthy human gut microbiome is stationary, seasonal, and non-random. Moreover, we demonstrate that it is self-explanatory to some extent, and its behavior can be predicted. The analysis of individual bacterial species uncovered the existence of three distinct longitudinal regimes in the healthy human gut microbiome. These regimes consist of 1) stationary and highly prevalent bacteria that exhibit resistance to environmental changes; 2) volatile bacteria that exhibit dynamic reactions to external stimuli, causing their presence to fluctuate over time; and 3) white noise. Clustering analysis revealed the presence of taxonomically diverse bacterial groups that exhibit similar fluctuations over time. In conclusion, our study highlights the importance of longitudinal data and provides new insights into the dynamics of the healthy human gut microbiome. We offer clear guidelines for clinicians and statisticians who conduct longitudinal studies and develop models to predict the behavior of the gut microbiome over time.

Web based tool to present genetic variant quantitative statistics on the map of Poland

Wojciech Frohmberg[1] Damian Ćwikliński[1] Beata Wiśniewska[1] Karol Zawiślak[1]

[1] *Poznan University of Technology*

While operating with quantitative statistics we often feel the need of geographical clustering data. In case of genetic variants, geographical context of samples can be especially useful. We could consider here exploration of traits ancestry, analysis of potential propagation of the mutation or even simple likelihood of genetic disease occurrence in a given location.

In the grant Genetic Map of Poland (GMP) presenting genetic variant data in the map was one of its basic principles. There was several expectations on the search engine of the tool: - the filtering should be responsive - various types of filtering should be allowed e.g. by: * simple chromosome position range loci, * dbSNP entry name queries, * ClinVar custom query loci, * geographical location in two layers: voivodeship, district - user should be able to choose the form of results presentation from: * the bunch of different chart types * tabular representation styles

The tool was created using well-established React/Node.js/Docker technologies that favor further development.

The role of tandem repeats in bacterial functional amyloids

Alicja Nowakowska[1] Jakub Wojciechowski[1] Natalia Szulc[1] Małgorzata Kotulska[1]

[1] *Politechnika Wrocławska*

Repetitivity and modularity of proteins are two related notions incorporated into multiple evolutionary concepts. We discuss whether they may also be essential for functional amyloids. Amyloids are proteins that create very regular and usually highly insoluble fibrils, which are often associated with neurodegeneration. However, recent discoveries showed that amyloid structure of a protein could also be beneficial and desired, e.g., to promote cell adhesion. Functional amyloids are proteins which differ in their characteristics from pathological, so that the fibril formation could be more under control of an organism. We propose that repeats in the sequence could be regulating the aggregation propensity of these proteins. The inclusion of multiple symmetric interactions, due to the presence of the repeats, could be supporting and strengthening the desirable structural properties of functional amyloids. It was observed that tandem repeats in bacterial functional amyloids have a distinct characteristic. The pattern of repeats supports the appropriate level of fibril formation and better controllability of fibril stability. The repeats tend to be more imperfect, which attenuates excessive aggregation propensity. Their desired structure and function is also reinforced by their amino acid profile. In the study, we focused on bacterial functional amyloids, due to their importance in biofilm formation. We propose that similar mechanisms could be employed in other functional amyloids, which are designed by evolution to aggregate in a desirable manner, but not necessarily in pathological amyloids.

Understanding Cross-talk between gut bacteria proteome and the aggregation of human proteins.

Jakub W. Wojciechowski[1] Alicja Nowakowska[1] Małgorzata Kotulska[1]

[1] *Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, 50-370 Wrocław, Poland*

The human gut microbiome is a complex microbial community consisting of a multitude of microbial species. Recent studies showed that its composition can be directly related to the occurrence of a range of diseases including amyloid-related ones, such as Parkinson's and Alzheimer's diseases or type II diabetes. Despite extensive research on this topic, details of this relationship remain elusive. One of the possible explanations for this phenomenon is interactions between microbial proteins and metabolites with host cells. Especially, amyloidogenic proteins produced by bacteria and fungi can shed light on this process. It is known that the onsets of Type II Diabetes, Alzheimer's, and Parkinson's diseases are closely linked to the pathological aggregation of proteins that form amyloid fibers. Interestingly, similar structures are produced by several microbial species to perform a wide range of vital functions from biofilm formation to molecular signaling. In this work, we aimed at identifying novel functional amyloids produced by the human gut microbiome and used a novel amyloid interaction prediction model PACT to predict their potential interactions with human disease-related proteins. Using a similarity search we identify a set of candidate functional amyloids and computationally showed that many of them are likely to interact with human proteins including amyloid beta, alpha-synuclein, or p53.

The exploration of Single Nucleotide Polymorphisms density in swine genome

M. Biłyk[1] M. Mielczarek[2] B. Nowak[3] J. Szyda[4] M. Fraszczak[5]

[1] *1 Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland*

[2] *2 Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland; Institute of Animal Breeding, Krakowska 1, 32-083 Balice, Poland*

[3] *3 Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland*

[4] *4 Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland; Institute of Animal Breeding, Krakowska 1, 32-083 Balice, Poland*

[5] *5 Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland*

Single Nucleotide Polymorphisms (SNPs) are valuable variants for identifying and localizing disease susceptibility genes or for understanding the molecular mechanisms of mutation. The aim of this study, was to analyse the density of SNPs in the swine genome. The dataset consisted of whole-genome DNA sequences of 12 Polish Large White pigs obtained using Illumina HiSeq2000 Next Generation Sequencing platform . There were 24,587,064 SNPs identified in the analysed genome. The first step, prior to conducting the analyses, was data preparation. After filtering and cleaning the data, distances between neighbouring polymorphisms were calculated. The analysis began with testing the normality of the distribution of inter-SNP distances. None of the tests confirmed the established null hypotheses. The homogeneity of the distance distributions was then examined, and it was determined between which chromosomes there were statistically significant differences. A second approach in examining the density of SNPs, was to divide each chromosome into 1000bp subregions to determine the most and least dense regions. We can also check whether the SNPs are uniformly distributed along the chromosome. The final step in analysing the density of polymorphisms, was to perform genomic annotation of the determined regions in the VEP program. After filtering data, 20,965,990 SNPs were considered for further analysis. The highest density of SNPs was on chromosome 10, and the lowest on chromosome 18. The distances between neighbouring SNPs are not uniformly distributed along the genome and their distribution depends on the chromosome. 46% of the SNPs located in the most dense regions were in intergenic regions and only 1% in coding regions.

Magnetstein: tool for complex mixture analysis in NMR spectroscopy

Barbara Domżał[1] Ewa Nawrocka[2] Dariusz Gołowicz[3] Michał Aleksander Ciach[1] Błażej Miasojedow[1] Krzysztof Kazimierczuk[2] Anna Gambin[1]

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

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

[3] *Institute of Physical Chemistry, Polish Academy of Sciences, Warsaw, Poland*

Modern metabolomics could not have existed without spectroscopy. Among its various types, nuclear magnetic resonance spectroscopy (NMR) stands out as reliable, non-destructive and non-invasive technology perfectly suited for the needs of contemporary metabolomics research. One of the problems in quantitative metabolomics is to determine relative abundances of chemicals present in a sample given its NMR spectrum. This problem can be addressed in various different ways, however, the approaches developed so far sometimes fail to accurately deconvolve mixtures that consist of strongly overlapping spectra of several components. We present a novel algorithm called magnetstein. The algorithm takes advantage from the usage of the Wasserstein distance, a metric known from fields such as probability theory, statistics and machine learning. The special properties of the Wasserstein distance together with our adjustments make our algorithm robust to unexpected shifts of peak positions, overlapping of spectra and presence of noise in the data. Moreover, magnetstein is outstandingly universal: it can be applied to spectra with different resolutions and disturbed lineshapes. Magnetstein is user-friendly, as it is a one-parameter method not requiring spectra's alignment nor extensive preprocessing, such as binning. The algorithm has been tested on several difficult datasets. We compared its performance with ACD/Spectrus 2020.1.1 and Mestrelab's MANIQ. In most of the cases, magnetstein beats the other tools distinctly. The algorithm is available in open-source Python package (see QR code).

Trans-homologous interactions identified in Hi-C data reveal putative gene regulation process

Magdalena Machnicka[1] Aleksander Jankowski[1]

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

The in situ Hi-C method combines the original Hi-C protocol with nuclear ligation, ensuring that interactions under study are located within the nucleus of a single cell. When performed on heterozygous cells, it allows for identification of interactions between chromosomes which form homologous pairs. Such interactions are known to be prevalent in fruit fly Drosophila melanogaster. Homologous chromosomes in this species tend to be tightly interacting and this phenomenon is known as homologous chromosome pairing. Paired homologous chromosomes exhibit high interaction frequencies along their entire length, but function of this global pairing is not well understood. For a collection of defined pairs of genetic loci, contacts between homologous chromosomes (trans-homologous contacts) have been shown to rescue regulatory interactions in cases when one of the haplotypes is deficient in its regulatory function (phenomenon known as transvection). Moreover, pairing is less pronounced in embryonic cells than in mature differentiated cells, suggesting that the establishment of trans-homologous interactions takes place during cell differentiation and might play a role in embryonic development.

Here we analyzed published in situ Hi-C data for heterozygotic D. melanogaster cells: embryos and a fully differentiated cell line. The Hi-C reads were phased, which allowed us to identify both cis- and trans-homologous interactions. We assessed the frequency of cis- and trans-homologous interactions for equally sized genomic bins and detected pairs of bins which exhibit significantly different frequency of trans-homologous interactions compared to cis-homologous interactions. In embryonic data we detected 3,450 such bin pairs, while for the cell line only 323, which is in agreement with expected increase in pairing in differentiated cells. In both cases, approx. 10

Cholesterol metabolism pathways disturbances in atherosclerosis - analyzes using stochastic Petri net-based model

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

[1]1 Institute of Computing Science, Poznan University of Technology, 60-695 Poznan, Poland

[2]2 Institute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland

[3]3 Department of Medical Chemistry and Laboratory Medicine, Poznan University of Medical Sciences, 60-806 Poznan, Poland

Atherosclerosis is a complex, global health problem causing significant morbidity and mortality worldwide. Despite decades of research, a complete cure for this disease remains elusive. Disturbances in cholesterol metabolism, local low-grade inflammation, and oxidative stress contribute to the development of atherosclerotic plaques. In this study, a stochastic Petri net model was constructed and analyzed to explore the impact of these factors on atherosclerosis. Through knockout and simulation-based analysis, we comprehensively investigated the underlying phenomena. Our findings indicate that solely blocking cholesterol's impact is insufficient to halt disease progression. Inhibiting oxidative stress alongside targets like PPI-1, MTTP, and HMGCR reduces foam cell accumulation, further supporting the efficacy of combined treatments against atherosclerosis.

Sliced Wasserstein distance as a similarity score for LC-MS data

Justyna Król[1] Michał Startek[1,2] Stanisław Grodzki[1] Anna Gambin[1]

[1] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland
[2] Institute for Immunology, University Medical Center of the Johannes-Gutenberg University Mainz, Mainz, Germany

Liquid chromatography–mass spectrometry (LC–MS) is an analytical chemistry technique used to identify chemical composition of complex mixtures. A very important part of many analyses of such data is comparing observed spectra with each other. Finding a suitable similarity score is a crucial part of that process. However, commonly used scores have many weaknesses. The Wasserstein distance (also known as the Earth's Mover distance) is known to be an effective metric for comparing one-dimensional MS spectra. However, calculating two-dimensional Wasserstein distance is computationally expensive, as it doesn't have a closed form solution that we get in the one-dimensional case. In order to reduce computational cost, we propose using Sliced Wasserstein Distance to calculate the distance between two LC-MS spectra. Radon transform is used to obtain a family of one dimensional linear projections of each spectra. Additionally, a modified stochastic gradient descent scheme is used to reduce a number of projections without a significant loss in quality of approximation. Another key part of analyzing LC-MS data is searching through spectra libraries in order to find the best matches to observed spectra. A library searching algorithm based on Sliced Wasserstein distance is proposed and examined. The algorithm uses partial evaluation with selection of best candidates through a priority queue, and iteratively expands most promising ones, enabling an orders-of-magnitude speedup over naive approach.

High-throughput transcriptomes analysis pipeline as a key tool to understanding the metamorphosis evolution in insects

Gabriela Machaj[1] Guillem Ylla[1]

[1]Laboratory of Bioinformatics and Genome Biology, Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University, Kraków 30-387, Poland

Insect metamorphosis is classified into hemimetaboly and holometaboly. Hemimetabolan nymphs are morphologically similar to the adult, whereas the holometabolans embryogenesis produces a larva dramatically different from the adult. The gene E93 was known to trigger adult morphogenesis in all insects in the last juvenile stage. We discovered that E93 is also necessary for the embryogenesis of the hemimetabolous Blattella germanica but is absent in the embryo of the holometabolous fly Drosophila melanogaster. This led us to the hypothesis that hemimetabolous insects display an adultiform body plan due to the embryonic expression of E93 which is absent in holometabolous. We developed a method to measure the E93 levels across dozens of insect species in embryos and pre-adult stages to test such a hypothesis. In our pipeline, we first queried NCBI's SRA database to identify embryonic and pre-adult transcriptomic datasets of insects and download their metadata. Using NCBI-datasets we retrieved a list of insects with the annotated genomes. By intersecting the metadata information of the transcriptomic datasets with the list of insects with annotated genomes, we identified 848 RNA-seq datasets belonging to 33 species for which we had RNA-seq data from both, embryo stages and pre-adult stages, and an assembled and annotated genome. RNA-seq datasets were downloaded and decompressed with SRA-Toolkit, cleaned with TrimGalore, and mapped to the corresponding genomes with STAR. All steps were automatized using Python scripts. The identification of E93 gene in each species was based on sequence similarity to that of fruit fly using BLAST. Next, featureCounts for R were used to generate the table of counts for each species, normalized as TPMs. To determine whether E93 was expressed in the embryo of each species, we compared the E93 TPM values in the embryo with those of the pre-adult stage which is known to have an expression peak in insects. Our results demonstrated that E93 expression is high in the embryos of hemimetabolous, but very low in the embryos holometabolous. These results support our hypothesis that E93 determines the nymphal genetic program and that the loss of embryonic E93 expression was instrumental in the origin of the holometabolan metamorphosis. This pipeline showcases the power of bioinformatics to leverage publicly available datasets from dozens of species to test hypotheses at a large scale without the need to generate new expensive datasets.

Exploring the potential significance of social networks for research in ecology

Rafał Miłodrowski[1,2] Guillem Ylla[2]

[1] Jagiellonian University, Doctoral School of Exact and Natural Sciences
[2] Jagiellonian University, Faculty of Biochemistry, Biophysics and Biotechnology, Laboratory of Bioinformatics and Genome Biology

Social media platforms have become prominent channels for public expression, allowing researchers to gain valuable insights into various aspects of human behavior and social trends. In this project, we explore the potential usage of commonly used social networks to gain valuable information for ecology. As proof of concept, we use the Twitter social network with queries regarding four animal species with well-known and clearly differentiated behaviors. These are the mosquitoes, crickets, monarch butterflies, and cockroaches.

By leveraging the Twitter API, we retrieve tweets that contain information about the four animals for a period of 12 years together with their metadata, which includes the post timestamp, the geolocation at the time of posting if the user allowed it, and the location that the user declared in their profile. This allows us to link the tweets to their associated geographical coordinates, and when not available, we used the location of the user's profile as a proxy. This approach enabled us to identify temporal and spatial patterns for each of these animals over time.

For example, for mosquitoes, we observed one large peak of activity from April to October while for butterflies two peaks of activity: one around March-April and the other around September-October. Crickets, on the other hand, show the greatest increase in activity in late summer, while cockroaches have a less pronounced period of activity falling over a wider period of summer. Tweets for all the queried animals showed distinct temporal and spatial patterns, which coincide with their known periods of activity. Our results suggest that within the massive amount of data in social networks, there is information regarding ecological phenomena that can be leveraged.

The findings from this research have important implications for ecological research, opening a new way to acquire insights into biological phenomena using large sources of data. We show that by integrating geolocalization information and temporal data from posts on Twitter over time we can track migrations, spatial distributions, seasonal patterns, and changes in these patterns over time of years. Similar approaches can be applied in the future to data from other social networks.

QUBO formulation of Multiple Sequence Alignment problem

Katarzyna Nałecz-Charkiewicz[1]

[1] *Warsaw University of Technology*

Possible formulations of the Multiple Sequence Alignment (MSA) problem in the form of QUBO (Quadratic Unconstrained Binary Optimization) will be presented. MSA, which is one of the key issues in the field of biological sequence analysis (to identify sequence homology), is usually solved using classical computational methods. However, due to the combinatorial nature of this issue, the question arises whether it is possible to use the potential offered by the quantum computing paradigm to solve it, in particular quantum annealers - quantum computers of special purpose, specialized for optimization tasks. We will show how MSA can be presented in the form of QUBO, which - apart from Ising models - is the basic way of defining tasks to be solved on the quantum annealer. A comparison of various MSA formulations will be presented as QUBO, using diverse synthetic and real-world biological sequences. We take into account factors such as computational resource consumption, time complexity, qubit count, and solution quality. The strengths and limitations of each formulation are examined concerning their practical use for actual quantum annealers.

Analysisi of CNV distribution in swine genome.

M. Kaźmierczak[1] M. Mielczarek[2] B. Nowak[3] J. Szyda[4] M. Fraszczak[5]

[1]Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland
[2]Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland; Institute of Animal Breeding, Krakowska 1, 32-083 Balice, Poland
[3]Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland
[4]Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland; Institute of Animal Breeding, Krakowska 1, 32-083 Balice, Poland
[5]Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland

Copy number variation (CNV) defined as changes in the copy number (gain or loss) of large genomic segments can overlapped a lot of functional elements of genome and they are equally important for fully understanding the mechanism of inheritance. Several studies have reported that pig is a model organism for molecular biomedical model for human diseases. In this study we aimed to determine the prevalence of CNVs formed de novo in the offspring genomes. Especially, we focused on the analysis of CNV inheritance in full siblings. Additional purpose of this study was checking whether the linkage disequilibrium (LD) structure significant differs between regions overlapped by duplications and those without any structural variants.

The analysed dataset consisted of whole-genome DNA sequences of 8 pigs representing the Polish Large White breed, obtained using the Illumina HiSeq2000. The genome average coverage ranged from 12 to 21. The alignment to the Sscrofa11.1 reference genome was done using BWA-MEM software, then Picard and SAMtools packages were used for post-alignment processing and CNVs were detected with the CNVnator and the Pindel programs. The statistical inference was performed using Cochran Q test as well as multidimensional scaling and tests for fractions.

The total number of deletions per individual ranged between 410 and 752 and duplications between 294 and 447. The length of CNVs varied from 400 to 194,900 bp (average 3,939.44 ± 10,975.9) for deletions and from 400 to 561,400 bp (average 12,377.2 ± 20, 154.3) for duplications. Most of deletions (duplications) arise de novo is located on chromosome 3 (16). Additionally deletions formed de novo are shorter than inherited. In the case of chromosome 12 duplications are located in the same genomic regions. Full siblings are more similar each other in the structure of CNVs than to other individuals as well as offsprings are similar to parental individuals. In all individuals, de novo duplications were found to be responsible for:

detection of a chemical stimulus involved in odor perception, coding for the G protein-coupled receptor signalling pathway and organization of cellular components. The value of LD between SNPs in duplications is significantly higher than in regions of the genome where no CNV was detected.

Finding key interactions in cohesin-dockerin complex from thermophilic organism to enable its use in synthetic biology applications towards green recycling of polymers

Wojciech Warmuz[1]

[1] *University of Adam Mickiewicz, Faculty of Biology*

Synthetic biology has gained increasing popularity as a potential solution to address limited supplies and environmental changes. Lignocellulosic biomass holds promise as a renewable resource for the production of cellulosic ethanol, a viable alternative to conventional fuels. Microorganisms such as fungi and bacteria can be employed for this purpose. However, many suitable bacterial hosts lack an efficient extracellular depolymerizing apparatus, limiting their ability to decompose complex materials independently [1]. To overcome this challenge, certain bacteria employ cellulosomes, which are cohesin-dockerin complexes composed of scaffolding proteins that enable the attachment of multiple enzymes to the cell surface. To enhance efficiency and enable industrial-scale application, elevated temperatures and thermostable enzymes are advantageous for lignocellulosic biomass hydrolysis. To achieve tighter binding within complexes, proteins derived from thermophilic organisms, which exhibit enhanced stability under harsh conditions, can be utilized. Since there are very few stable cognate cohesin-dockerin complexes to provide sufficient binding strengths, we have investigated interactions in cohesin-dockerin pair from thermophilic organism Hungateiclostridium clariflavum (Hc) that exhibit intriguing cross-interactions with the pair from Acetivibrio cellulolyticus (Ac), a natural producer of cellulosomes. Due to the lack of crystal structures for Hc proteins, Alphafold2 model was used as input. Using molecular dynamics (MD) simulations, ensembles of both pairs were generated to elucidate key interactions governing their stability and cross-reactivity using MM-GBSA energy decomposition. The conformations of these complexes obtained from MD simulations were further used for ensemble-based computational saturation mutagenesis of non-critical interface residues to design more tighter binding pairs. This knowledge will enable effective experimental validation of critical interactions and guide future engineering of the complex. Given the limited understanding of bacterial structures and their inability to withstand harsh conditions present in during biomass processing, insights into the properties and behaviors of these new structures are crucial for advancing research related to cellulose recycling.

[1] Dvořák et al., 2020: Surface Display of Designer Protein Scaffolds on Genome-Reduced Strains of Pseudomonas putida

mONiTor: real-time monitoring of Oxford Nanopore sequencing run

Wiktor Kuśmirek[1]

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

Nanopore sequencing is the fourth-generation DNA sequencing technology and the significant advantages of nanopores include ultralong reads, low material requirement, and high throughput. Along with the development of the nanopore technology itself, open-source tools supporting work with the sequencers should also be developed.

Herein, we presented mONiTor - the new tool for monitoring the nanopore sequencing process. The tool monitors the metrics available in the sequencer, the state of the computer to which the sequencer is connected, and the contents of the fast5 and fastq files. The metrics are stored in the Prometheus database and presented in interactive diagrams using the Grafana software. Thanks to the technologies used, the user can easily log in and view the sequencing status (current and archived). In addition, after appropriate configuration, the user can be informed by e-mail about exceeding the limit, e.g. 90

NSRC-Search: Efficient searching for similar protein sequences of non-standard amino acid composition

Patryk Jarnot[1]

[1]*Department of Computer Networks and Systems, Silesian University of Technology*

Protein sequence fragments with non-standard amino acid compositions are currently of interest to scientists who are discovering the biological roles of proteins. Standard amino acid sequence consists of a highly diverse set of residues, but many proteins also contain fragments that are ordered, compositionally biased to some residues, or consist of only a few residue types. Similarity to proteins with known properties can be used to conclude about proteins of unknown properties. These conclusions are used to reduce the number of research scenarios of experimental methods, consequently reducing their time and cost. Several methods already exist to identify domains with non-standard amino acid composition, but we lack methods designed to compare them. Recent studies have also shown that canonical methods for protein sequence comparison, including BLAST, HHblist and CD-HIT are suboptimal when analysing low-complexity domains. Therefore, we propose the NSRC-Search method, which handles similar protein sequence fragments with unusual composition.

This method first uses existing protein domain identification methods (e.g. CAST, SEG or fLPS) to create a database. For each query sequence, the method calculates composition of residues, which is further used to recalculate a scoring matrix and filter out irrelevant domains. The scoring matrix is recalculated to promote frequently occurring residues found in the query domain. The method then computes Jaccard score for each database and query sequence. If the score is below a certain threshold, the sequence is filtered out. Otherwise, NSRC-Search uses the Needleman-Wunsch algorithm and the recalculated scoring matrix to align sequences. If the alignment score, similarity or identity between sequences is above a user specified threshold, then the alignment is reported. We compared this method to BLAST by analysing the similarity between domains identified by CAST in the UniProtKB/Swiss-Prot database. Our preliminary results show that NSRC-Search reports more similar sequences that are compositionally biased to the same residue.

Obtaining new knowledge from biomedical studies using robust aggregative feature selection (RAFS)

Radosław Piliszek[1] Witold Rudnicki[1,2]

[1] *Computational Centre, University of Bialystok, Bialystok, Poland*

[2] *Institute of Informatics, University of Bialystok, Bialystok, Poland*

Feature selection in data analysis is fundamental. There are many feature selection methods in the literature, with varying end goals. The approach presented here discovers a minimal-optimal feature set for binary classification without involving any particular machine-learning model. It is based on information-theoretic measures and designed to be robust by employing an internal cross-validation scheme and popularity ranking. We define robustness as avoidance/minimisation of bias/over-fitting and enhancement of stability as measured using Jaccard index and consistency score. We show how this approach excels in comparison to well-known methods, such as minimum redundancy maximum relevance (mRMR) and redundant feature elimination (RFE). Moreover, we present how it helped to discover new knowledge about carcinoma in situ (CIS) in the course of bladder cancer (BLCA), about kidney renal cell carcinoma (KIRC) and about the effect of Epstein-Barr virus (EBV) in the course of lymphoproliferative disorder (PTLD).

Comparison of Chromatin Structural Features in Human ESC-H1 Mapped Using GAM or Hi-C

Teresa Szczepińska[1,2,3] Christoph Thieme[1] Sachin Gadakh[2] Alexander Kukalev[1] Warren Winick-Ng[1] Rieke Kempfer[1] Thomas Sparks[1] Miao Yu[4] Bing Ren[4] Dariusz Plewczynski[2,5] Ana Pombo[1]

[1]*Berlin Institute for Medical Systems Biology, Max-Delbrück Centre for Molecular Medicine, Berlin 10115, Germany*

[2]*Center of New Technologies, University of Warsaw, Warsaw 02-097, Poland*

[3]*CEZAMAT, Warsaw University of Technology, Poleczki 19, 02-822 Warsaw, Poland*

[4]*Ludwig Institute for Cancer Research, La Jolla CA 92093, USA*

[5]*Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw 00-661, Poland*

Understanding the limitations of chromatin mapping techniques in detecting biological aspects of genome 3D structure is important for discovering its function, efficiently assessing long-range gene regulation, and making comparisons between cell types. Whereas Hi-C contact maps are based on the frequency of proximity ligation in millions of cells, Genome Architecture Mapping (GAM) extracts spatial information about 3D genome topology by sequencing the genomic content of hundreds to thousands ultra-thin, randomly oriented nuclear slices. We have performed GAM on H1 human embryonic stem cells and compared GAM data with Hi-C data from the same cell line available on 4D Nucleome repository. We have measured the detectability of each method along the linear genome at 50 kb genomic resolution. We devised methods to robustly remove regions with detectability outliers in either method. Moreover, we have characterized genomic windows according to the presence of functional features such as histone modifications, protein binding, gene density and expression (GRO-seq, RNA-seq), chromatin accessibility (ATAC-seq), enhancers and lamina association. Our analyses indicate that the two methods have good similarity at TAD level but show differences in the assigned compartments. We also observed different capabilities in detecting functionally active genomic regions. While more windows need to be cleared out from Hi-C than from GAM because of low detectability, we noticed that regions annotated with features are more often removed from one dataset exclusively.

Day 2 - 14 September 2023

Keynote speaker - the EMBO Lecture

Knot or not? Sequence-based identification of knotted proteins with machine learning

Joanna Sułkowska[1]

[1] *CeNT Centre of New Technologies, University of Warsaw, Poland*

Knotted proteins, although scarce, are crucial structural components of certain protein families, and their roles remain a topic of intense research. Capitalizing on the vast collection of protein structure predictions offered by AlphaFold, this study computationally examines the entire UniProt database to create a robust dataset of knotted and unknotted proteins. Utilizing this dataset, we develop a machine learning model capable of accurately predicting the presence of knots in protein structures solely from their amino acid sequences, with our best-performing model demonstrating a 98.5% overall accuracy. Unveiling the sequence factors that contribute to knot formation, we discover that proteins predicted to be unknotted from known knotted families are typically non-functional fragments missing a significant portion of the knot core. The study further explores the significance of the substrate binding site in knot formation, particularly within the SPOUT protein family. Our findings spotlight the potential of machine learning in enhancing our understanding of protein topology and propose further investigation into the role of knotted structures across other protein families.

Session 3

Regulatory mechanisms in the RNA World based on short RNA sequences

Jarosław Synak[1,2,3] Agnieszka Rybarczyk[1,2,3] Jacek Błażewicz[1,2,3]

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

[2] *European Center for Bioinformatics and Genomics*

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

Modern cells have an impressive amount of different mechanisms controlling nearly all processes which take place inside them. Some of these however, like riboswitches or cyclonucleotides, appear to be much older than the rest, possibly predating DNA. There is a possibility that they had evolved from ancient mechanisms in the RNA World, which helped the first biological molecule populations to survive. The authors propose a theory, which could explain how a simple control mechanism could have formed and how efficient it actually could have been. The core of it are short RNA molecules which could act as inhibitors, binding to the corresponding RNA sequences and preventing them from being replicated. It would work based on Watson-Crick pairing and chemical reaction equilibrium. Inhibitors would also bind to other very short molecules – anti-inhibitors, a reaction which would help to stabilise the concentration of inhibitors themselves. The model was thoroughly analysed using both theoretical and simulational methods. Three different algorithms were developed in order to check different assumptions. The most important result was that such system in theory could work, blocking surplus molecules and releasing them whenever the population started to dwindle. This mechanism would increase the chance of survival and further evolution of an RNA population, especially in later stages when different RNA enzymes had probably evolved and had to be kept in equilibrium, just like enzymes in modern cells. The inhibitors themselves could have been controlled by natural selection of entire populations – groups of molecules with non-functional control mechanisms were at a disadvantage and ended up being dominated.

Application of deep generative models for RNA 3D structure prediction.

Marek Justyna[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*

The adaptation of deep learning (DL) models was a breakthrough in many fields of science, including molecular biology. In the latter, the most famous example is AlphaFold, the system for protein structure prediction that won the 2022 Breakthrough Prize in Life Sciences. The potential of deep learning strongly depends on the size and diversity of the training datasets. In the field of RNA 3D structure prediction, the amount of high-resolution structures still remains low in contrast to proteins, so the results of DL-based models for RNA are not as spectacular as in the case of the AlphaFold. The alternative approach to dealing with a small amount of data is to apply deep generative models (DGM), such as generative adversarial networks (GAN) or denoising diffusion probabilistic models (DDPM). In the presentation, we will discuss the potential of generative models in the prediction of 3D RNA structure.

COMA - novel tool for aligning optical mapping data

Norbert Dojer[1] Mikołaj Arciszewski[1] Zofia Kochańska[1]

[1] *University of Warsaw*

Advancements in DNA sequencing technologies have revolutionized the field of genomics, enabling researchers to generate vast amounts of genomic data at an unprecedented scale. Among these technologies, optical mapping has emerged as a promising approach for long-range genome analysis. It leverages the power of fluorescence microscopy to directly visualize and map DNA molecules, providing valuable insights into structural variations and genome organization.

Accurate alignment of optical mapping sequences is crucial for interpreting and extracting meaningful biological information. However, due to the unique characteristics of optical maps traditional sequence alignment approaches often fall short in providing accurate alignment results. To address this challenge, we present a novel tool called COMA (Cross-correlation Optical Map Alignment), specifically developed for aligning optical mapping sequences and overcoming these difficulties.

COMA tackles this problem by incorporating a double cross-correlation approach that leverages the specific features of optical mapping data. It utilizes two separate computations of cross-correlation to first identify potential locations where a molecule could be mapped and then perfects the mapping. The tool also incorporates an extensive set of parameters that can be adjusted to fit the currently studied data and modify its stringency. What makes it even more unique is the additional output of locations where there were noticeable conflicts in alignment. These discrepancies can provide additional information on potential genomic regions where structural variants can be observed.

Additionally, we investigate the impact of COMA on downstream analysis tasks, such as structural variant detection. We compare our results with those obtained by existing tools. We have created a novel tool which identifies structural changes and verifies them against the existing benchmark dataset, which serves as a gold standard. It is adapted to use a file format that is standard for aligned optical maps, allowing it to be used on results obtained using different tools.

In conclusion, this study presents COMA, a novel tool specifically designed for aligning optical mapping sequences. We believe that COMA will serve as a valuable resource for researchers working in the field of genomics, offering new opportunities to gain deeper insight into the structure of genomes.

Session 4

Microbiome health - it is time to redefine it?

Kinga Zielińska[1,2] Dagmara Błaszczyk[1,3] Krzysztof Mnich[4] Witold Wydmański[1,3] Valentyn Bezshapkin[1,5] Tomasz Kosciolek[1] Witold Rudnicki[4,6] Paweł P. Łabaj[1]

[1] *Jagiellonian University, Malopolska Centre of Biotechnology, Krakow, Gronostajowa 7a*
[2] *University of Oxford, Nuffield Department of Clinical Medicine, Old Road Oxford OX3 7BN, United Kingdom*
[3] *Jagiellonian University, Doctoral School of Exact and Natural Sciences, Krakow, Lojasiewicza 11*
[4] *University of Białystok, Computational Centre, Bialystok, Konstantego Ciołkowskiego 1M*
[5] *ETH Zurich, Institute of Microbiology, Zurich, Vladimir-Prelog-Weg 4*
[6] *University of Białystok, Institute of Computer Science, Bialystok, K. Ciołkowskiego 1M*

Ability to evaluate one's health status based on a gut sample, independently of clinical diagnosis, has been a focus of numerous human microbiome studies. Current approaches are restricted to measuring alterations in the "core microbiome", a set of taxa most frequent among healthy individuals, as potential signs of dysbiosis. Our analyses of the Human Microbiome Project2 (HMP2) samples, however, identified functional aspects of the gut microbiome as more conserved among healthy individuals. This aligns well with previous hypotheses that a functional profile can be achieved by various combinations of taxa and their interactions, therefore, a shift towards defining health status based on the functional profiles is required. In our research, we analyze a wide range of healthy and dysbiotic samples from the Human Microbiome Project and American Gut Project. We choose MultiDimensional Feature Selection algorithm (MDFS) as our feature selector to identify key species that, when occurring together, improve the separation of samples from different groups. This unique approach leads to a better understanding of the functional profiles of the microbiome by identifying interactions of species that play the greatest part in maintaining it. In healthy samples, species form symbiotic relationships with one another to provide metabolic products necessary for the survival of the whole ecosystem. In the advent of dysbiosis, those interactions are distorted and alternatives, usually in the form of new interactions or rare functions, are found. The identification of "healthy" microbiome behaviors is key to understanding its homeostatic state, which has important applications in further human health studies. Therefore, our ultimate goal is to better understand the microbial synergies that influence the functional aspects of the microbiome and, ultimately, to redefine what is currently known about human gut microbiome health.

Encoding and decoding functional information of Low Complexity Regions in word embeddings vectors

Sylwia Szymanska[1] Aleksandra Gruca[1]

[1] Department of Computer Networks and Systems, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland

Low Complexity Regions (LCRs) are fragments of protein sequences characterized by low amino acid diversity. The peculiar appearance of LCR sequences has contributed to opinions that they are non-functional fragments and for a long time were ignored by scientists. Currently, LCRs are extensively studied. Unfortunately this knowledge has not been systematized yet and the only proven source of information about their functionality are publications. Manual extraction of such information from text is time consuming and sensitive to human error. To address this problem, we applied language models to analyze publications containing information about LCRs using as a case study the following functions: nucleic acids binding, phase separation and aggregation. Using dimensionality reduction method we created one representative embedding vector for each publication. However correct classification of vectors created on token level is challenging, because similar tokens have close embedding values. To check if our approach is not prone to this problem, we added so-called hard cases - publications that were unrelated to LCRs but could not be simply removed by filtering for keywords. We manually annotated 8980 publications: 3493 positive publications that describe LCR and its function, and 5487 negative ones including 680 hard cases. Those scientific texts were split into training and test dataset. To show that our language model is able to correctly encode the information on LCR function in a single vector text embedding, we performed classifications using the random forest algorithm. To assess the quality of our approach, we compared the results with the baseline LitSuggest model. LitSuggest returns a list of ranked publications based on information extracted from the text. Using the same data to train and test both models, our model had a precision of 93.16 and F1 score of 92.87, while LitSuggest had 82.45 and 87.85. The baseline model got a recall of 94.00, while our model's recall was 92.60. To decode information from a single embedding vector we check the similarity between representative vector and reference vector(s) created based on tokens describing specific LCR and function. We developed a language model that is capable of successfully classifying publications containing LCR and its function based on information extracted from the text. We also show that we are able to encode and decode information about LCRs and functions in a single vector of text embedding.

Co-occurrence of amino acids in low complexity regions in proteins

Joanna Ziemska-Legiecka[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 Network and Systems, Silesian University of Technology, Gliwice, 44-100, Poland*

Low complexity regions (LCRs) are regions in proteins characterized by a high density of only several types of amino acids. LCRs can be homorepeats, tandem repeats of short fragments, or just regions with fuzzy amino acid composition. For a long time, those regions were considered non-functional, but with further research, scientists discovered many functions in fragments. Additionally, we noticed that certain types of amino acids co-exist in sequences more often than others, so in this work, we decided to analyze this hypothesis. In order to check the co-occurrence of amino acids, we have made some statistical analysis on the dataset of LCRs from Uniref90. LCRs were identified with the SEG method with the usage of very restrictive parameters. It allowed us to find mostly repetitive regions, without false positive LCRs. In the next step, we divided LCRs into datasets. Each dataset is composed 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 10For obtained co-occurring pairs of amino acids in the test, we checked their scores in the BLOSUM62 and PAM30 matrices and compared their frequency to all scores in these matrices. As a result, we noticed that pairs have more positive scores from matrix PAM30 than all scores from this matrix. In the matrix BLOSUM62, our pairs have the same distribution as scores in a total matrix.

Sliced Wasserstein distance for creating precursor-fragment relationships in data-independent acquisition proteomics

Stanisław Grodzki[1] Michał Startek[1,2] Justyna Król[1] Anna Gambin[1]

[1] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Warsaw, Poland,
[2] Institute for Immunology, University Medical Center of the Johannes-Gutenberg University Mainz, Mainz, Germany

Liquid chromatography-mass spectrometry is a high-throughput experimental technique for identifying the chemical composition of complex mixtures. Data-independent acquisition (DIA) enables comprehensive and unbiased analysis of such samples. DIA method results in highly multiplexed data which require specialized tools for further downstream analysis. The Wasserstein distance is a well-known similarity measure for one-dimentional mass spectra. However, it is computationally expensive and does not have a closed form solution in the multi-dimentional case. We present a novel, more efficient approach using Sliced-Wasserstein distance to create precursor-fragment relationships for ions across the entire mass to charge range. Additionally, we use interval tree data structure to filter out non-matching spectra and improve computational complexity. We show that obtained approach spectra coming from LC-IMS-Q-TOF experiments display DDA quality.

Unseen passages – investigating transient tunnels facilitating ligand transport in dehalogenase.

Igor Marchlewski[1,2] Jan Brezovsky[1,2]

[1] *International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland*
[2] *Laboratory of Biomolecular Interactions and Transport, Department of Gene Expression, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland*

Although challenging to study, transport tunnels represent one of the major contributor to the efficiency and selectivity of enzymes with buried active sites. Tunnels leading to the active site exhibit different geometrical and chemical properties discriminating passage of various compounds into/from active site including substrates, products, or solvent. Tunnels are dynamic entities, hence their properties are constantly changing. Up to date, most studies of transport tunnels focused mainly on the well-defined tunnels, whereas the role of transient tunnels in the ligand transport and specificity is understudied. Due to their transient nature, the opening events might not be observed in the timescale of conventional molecular dynamics (MD) simulation. Therefore, in this study we applied state-of-the-art enhanced sampling method, Gaussian Accelerated MD, to boost the occurrence of rare events such as transient tunnels opening. Here, we selected Dehalogenase DhaA as a model system due to its well-established tunnel network and availability of ample experimental data [1,2]. Putatively relevant tunnels identified from enhanced simulations were for docking of substrate, and product molecules, followed by adaptive sampling high-throughput MD to generate series of simulations describing the transport events via all tunnels. Finally, Markov State Modeling was applied to define probability of ligands in meta-stable states and interconversion between them. These models describing transport of ligands in along each identified pathways was used to obtain estimates of transport rates and potential obstacles in transition of ligands between the active site and bulk solvent. Our methodology provides quantitative description of transport and sheds light on the importance of transient tunnels for protein function. As a result, it constitutes a reliable starting point for rational engineering aiming at designed enzymes with improved transport to/from the active site or modulated selectivity by adjusting discriminating properties of the tunnels. This work was supported by the National Science Centre, Poland (grant no. 2017/26/E/NZ1/00548). The computations were performed at the Poznan Supercomputing and Networking Center.

1. Pavlova, M. et al. Nat Chem Biol 5, 727–733 (2009).
2. Klvana, M. et al. Journal of Molecular Biology 392, 1339–1356 (2009).

Keynote speaker

Unraveling the Microbial Enigma: Overcoming Bioinformatics Barriers in the Final Frontier

Kasthuri Venkateswaran[1]

[1] California Institute of Technology, Jet Propulsion Laboratory; Biotechnology and Planetary Protection Group, Pasadena, California, USA

We characterized the microbiomes of environmental surfaces and atmospheric samples within the International Space Station (ISS) to understand their relationship to crew health and hardware maintenance. Through the Microbial Tracking projects, we created a detailed microbial census of the ISS environments using both advanced molecular microbial community analyses and traditional culture-based methods. These "omics" methodologies yielded an extensive microbial census, providing significant insights into the changes in populations of beneficial and potentially harmful microbes induced by spaceflight.

We will discuss the lessons learned from ISS missions about microbial prevalence using iTag sequencing, metagenomics, and resistomes. Additionally, while characterizing approximately 3,000 bacterial and fungal strains, we discovered several novel species, and we will present the characterization of these new species. Our research also revealed the virulence characteristics of fungi and the production of secondary metabolites of biotechnological importance.

The findings from the Environmental "Omics" project (basic science) should be leveraged to enhance human health and well-being within closed systems. In other words, the goal of the microbial tracking research is to "translate" findings from fundamental research into medical practice (pathogen detection) and meaningful health outcomes (countermeasure development). The "omics" data sets have been added to the NASA GeneLab bioinformatics environment, which includes a database, computational tools, and improved methods. This will be made open to the scientific research community to foster innovation.

Day 3 - 15 September 2023

Keynote speaker - Honorary PTBI Member

Influence of passenger mutations on expansion and extinction of cancer clones

Andrzej Polański[1]

[1]Silesian University of Technology, Gliwice, Poland

Tumour evolution is strongly related to somatic mutations occurring in evolution of cancer cells. Somatic mutations in cancer cells are classified as either driver or passenger. Driver mutations occur very rarely in cellular replications, but have strong causative effect on cancer development. Majority of somatic mutations found in DNA are passenger mutations, which do not exert strong effect on the tumour growth. Initially passenger mutations were considered as fully evolutionary neutral. However, numerous recent experimental studies provide evidence that passenger somatic mutations are likely to exert either weakly deleterious or weakly advantageous effect on tumour expansion.

In the talk experimental evidence for evidence that accumulated passenger mutations can make an impact on cancer evolution, parallelly to drivers, is first reviewed. Arguments include (i) existence of molecular signatures distinguishing between cancer types and their progression scenarios (ii) the use of molecular functional impact scores, which prove that aggregated effect of passengers plays role in tumorigenesis.

Next, results of stochastic simulations (with Gillespie algorithm) of the model of cancer cells population evolution driven by rare, strongly advantageous driver mutations accompanied by frequent passenger mutations, which are either mildly advantageous or mildly deleterious, are shown. In simulations, cancer cells population stratifies into subpopulations defined by composition of driver mutations called cancer clones. In the course of evolution of cancer clones the phenomenon of clonal interference is observed. Clonal interference consists of events of emergence of new clones and extinction of old ones. Emergence and extinction of clones is under the influence of aggregated effects of passenger mutations, which occur in the evolution. The influence of cumulated passenger mutation of two types is quantified and compared to experimental data.

Session 5 - PTBI Laureates

Development of numerical tools for screening biologically active compounds for antimicrobial effects

Mateusz Rzycki[1]

[1] *Department of Biomedical Engineering, Wroclaw University of Science and Technology*

The spread of antimicrobial resistance was identified by the WHO as one of the top 10 global threats to human health. Unless new drugs are discovered, millions of people will fall victim to drug-resistant bacteria (DRM) each year. One of promising approach is to focus on antiseptics that do not have a well-defined molecular target in bacterial cells. The cationic surfactants can selectively attack cell membranes inducing their destruction by emulsification. Their specific mode of action makes them promising in the fight against DRM. The main aim of the work was to develop numerical approaches for rapid assessment of the antimicrobial activity of cationic agents.

The initial phase was focused on molecular interactions of commercially available agents such as octenidine (OCT) on various membrane models. I investigated the agent's behavior and described the changes in significant membrane parameters. The in-depth analysis allowed to propose a novel mechanism of selective OCT activity based on differences in mechanical properties between bacterial and eukaryotic bilayers.

Sequentially, other promising agents with strong antimicrobial activity were investigated. I collected a database, characterized 250 compounds and performed quantum calculations for their force field optimization. Then, proposed a standardized protocol based on molecular dynamics simulations and tested 25 molecules on E.coli membrane model. Analysis of bilayer parameters allowed to select the 8 promising precursors with indication for the strongest antimicrobial activity.

In the end, I proposed a novel software called Diptool. Simplifying the interaction scheme allowed for the development of a tool for screening membrane-drug interactions. The major difference from other methods is that Diptool is a targeted solution based on dipole interactions. Significant parameters (selected with QSAR) such as dipole moments, partition coefficient and agent size were implemented in the calculation core. This resulted in a fast and effective screening of free energy profiles with a great speedup compared to standard approaches. The analysis of energy distribution allows to estimate the potential antimicrobial activity of the drug.

Understanding the mechanism of action of many agents is possible through multiple approaches. The implementation of various molecular insights into the drug's activity could open a new path for the development of more effective antibacterial compounds against DRM

GrassSV – hybrid method to detect structural variant in high throughput DNA-seq data

Dominik Witczak[1]

[1] *Poznan University of Technology*

Every living organism contains instructions in the form of genetic code, which determines its appearance and function. This code is stored in genes, which encode functional proteins. Even a minor mutation may change the protein structure altering it's function. One possible example of this is sickle cell anemia which is caused by a mutation affecting the structure of globin (a protein, component of hemoglobin). Furthermore mutations outside the gene sequence may also include regulatory elements affecting genes expression.

Most often, changes in the genome sequence are detected based on a reference genome, which represents the genome of an individual from a particular species. The sample genome is compared to the reference genome, and a list of differences is determined. The differences (mutations) are categorized into size based groups. The shortest changes are single nucleotide polymorphisms (SNPs). Longer changes, up to 1000 nucleotides, are called insertions and deletions (INDELs), while the longest ones are structural variants (SVs). Regardless of their length, each mutation can contribute to the development of a disease or drug resistance. Often, it is presence of multiple mutations combined with environmental factors that lead to a specific diesease manifestation. In this study, the focus is on detecting structural variants.

The GrassSV pipeline utilizes a comprehensive, hybrid pattern matching approach to detect various types of structural variants (SVs) in a single run. By leveraging the fact that the majority of the genetic code is conserved across individuals, GrassSV performs depth of coverage analysis and contig assembly to identify potential sites of genetic variation and extract reads with high informational value. This efficient approach significantly reduces the cost and computation time required for DNA assembly while allowing for precise detection of SV types and breakpoint locations.

RNApdbee 3.0: webserver for 3D RNA structure analysis

Kamil Niżnik[1] Paweł Śnioszek[1] Gabriel Wachowski[1] Mikołaj Żurawski[1]

[1] *Poznan University of Technology*

Many researchers in the field of RNA structural biology and bioinformatics find access to correctly annotated RNA structure to be a topic of major importance, especially in the secondary and tertiary structure predictions. RNApdbee webserver, introduced in 2014 and followed by RNApdbee 2.0 in 2018, primarily aimed at the problem of secondary extraction from PDB or PDBx/mmCIF format using multiple heuristic algorithms with option to visualize the extracted structures using multiple tools. It allowed analysis of isolated base pairs' impact on RNA structure. It could visualize RNA secondary structures-including that of quadruplexes-with depiction of non-canonical interactions. It also annotated motifs to ease identification of stems, loops and single-stranded fragments. RNApdbee 3.0 introduces multiple breaking changes compared to its predecessors while keeping the same functionality. It terminates usage of heuristic algorithms for pseudoknot assignment problem in favor of an optimal approach, thanks to integration with Gurobi Optimization tool. Next advancement is robust integration with tools for base pair analysis and secondary structures' visualization that were not available in the 2.0 version. On top of that, it provides modern Graphical User Interface.

Deep neural autoencoder tool and unsupervised learning for scRNA-Seq data exploration

Anna Mrukwa[1,2] Joanna Zyla[1]

[1]Silesian University of Technology, Department of Data Science and Engineering, Akademicka 16, 44-100, Gliwice, Poland
[2]Cufix, 05-825 Grodzisk Mazowiecki, Poland

An important step in the expansion of medical knowledge was the development of genome sequencing techniques, with scRNA-seq taking a special place due to its success in capturing the activity across all genes for a single cell. Usage of these tools enables us to research the heterogeneity of the tissues and cells constructing them, allowing the recognition and insight into the malignant lesions. To meet the need to properly investigate such cell properties, both classical and deep Machine Learning algorithms are commonly used. In this work, the SAUCIE model is reimplemented and compared to other popular methods. The results are measured in terms of the quality of dimensionality reduction and clustering. SAUCIE consists of a cascade of modified autoencoders when batch correction is needed and only one otherwise. The embedding creation is a typical task for this kind of model, thus by forcing the latent space to be 2D, the dimensionality reduction for data visualization can be achieved. For the batch correction, besides the reconstruction task of AE which focuses on the reference batch, a dedicated additional loss function appears in the first part of the cascade: MMD loss strives to reduce the technical artifacts without removing the biological differences. For clustering, a special layer with near binary activations via the usage of ID regularization was introduced. To make the groups more concise, intracluster distances are calculated and minimized for the binary codes. To counter this effect and minimize the number of clusters, ID regularization loss is implemented via the usage of von Neumann entropy, forcing the neural network to spread the activations more sparsely and unevenly. For the comparison of the clustering, classical methods such as k-means, hierarchical clustering and Louvain were used. Additionally, a new k-means based method, DiviK, was evaluated. For dimensionality reduction, popular algorithm PCA was used alongside t-SNE and UMAP ran on the reduced space generated via PCA. The model proved to be influenced by the drawbacks introduced with the autoencoder usage - the probable latent space collapse. A possible solution to this issue would be to implement a VAE instead. What could also be seen was the significant superiority of the Louvain algorithm for the clustering. Thus, the possible further research on the graph algorithms, even their neural network variation, could prove to produce interesting results for single-cell

clustering.

Keynote speaker

Molecular representation learning for drug discovery

Djork-Arné Clevert[1]

[1] *Pfizer, Berlin, Germany*

Recently, molecular representation learning has solidified its position as an indispensable instrument in chemoinformatics, driving significant strides in drug discovery by enhancing the precision of bioactivity and toxicology predictions. In this presentation, I will delineate the foundational principles of molecular representation learning. Further, I will elucidate their application in bioactivity prediction and biological-conditioned drug de novo design. I will conclude by emphasizing their pivotal role in addressing molecular inverse problems.

Sponsors

华大基因

BGI

Ministry of Education and Science
Republic of Poland

Projekt dofinansowany ze środków budżetu państwa, przyznanych przez Ministra Edukacji i Nauki w ramach Programu Doskonała Nauka II (nr rej. KONF/SP/0441/2023/01)

Politechnika
Śląska

UCZELNIA
BADAWCZA
INICJATYWA DOSKONAŁOŚCI

Organizing Committee

- Aleksandra Gruca, Department of Computer Networks and Systems, SUT, Gliwice (chair)
- Paweł Łabaj, Małopolska Centre of Biotechnology, UJ, Kraków
- Joanna Żyła, Department of Data Science and Engineering, SUT, Gliwice
- Tomasz Kościółek, Department of Data Science and Engineering, SUT, Gliwice
- Patryk Jarnot, Department of Computer Networks and Systems, SUT, Gliwice
- Justyna Mika, Department of Data Science and Engineering, SUT, Gliwice
- Sylwia Szymańska, Department of Computer Networks and Systems, SUT, Gliwice
- Joanna Tobiasz, Department of Data Science and Engineering, SUT, Gliwice

Author Index