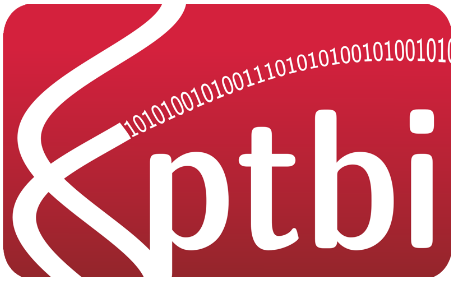
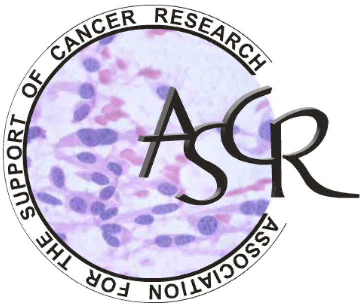
BioMeeting 2024

Kroczyce, May 24-26 2024

Partners of BioMeeting 2024





Organizing Committee

Silesian University of Technology (main organizer):

Kamila Szumała
Alicja Bargieła
Urszula Cojg
Patrycja Podleśny
Emilia Witek-Żelazny

Wrocław University of Science and Technology:

Nikola Rybarczyk
Adam Gruda

Jagiellonian University:

Sylwia Bożek

University of Warsaw:

Joanna Dąbrowska
Aleksander Janowiak

Scientific Committee

Silesian University of Technology:

Dr hab. inż. Aleksandra Gruca, prof. PolŚl (Committee chairwoman)
Dr inż. Joanna Żyła
Dr inż. Patryk Jarnot

Wrocław University of Science and Technology:

Prof. dr hab. inż. Małgorzata Kotulska
Dr hab. inż. Sebastian Kraszewski, prof. PWr
Dr inż. Marlena Gąsior-Głogowska
Mgr inż. Beata Borysiuk

University of Warsaw:

Dr hab. Bartosz Wilczyński, prof. UW (Committee chairman)
Dr Aleksander Jankowski

Jagiellonian University:

Dr hab. Krzysztof Murzyn, prof. UJ
Dr hab. inż. Paweł Łabaj

Book of Abstracts

Presentation Abstracts 5

Application of long-read sequensing to improve genotyping of complex pharmacogenetic regions . . . 6

Clinical trials. Real World Data as a tool of implementation of innovative medical technologies. . . . 7

Development of WGS pipelines for computation and reporting of polygenic risk scores 8

iFlow platform for automated interpretation and customised reporting of NGS data in genome-based medicine . . . 9

Secrets of the link between neurodegeneration and bacterial amyloid migration 10

Poster Abstracts 11

A deep dive into invariant Natural Killer T (iNKT) cells a single-cell transcriptomics perspective . . . 12

A new method for protein function prediction . . . 13

Adopting the linear regression model to characterize the diversity profiles of TCR repertoires. . . . 14

Analysis of heterogeneity in hematopoietic stem cell differentiation trajectories using RNA sequencing at the single-cell level. . . . 15

Analysis of intracranial pressure in patients with suspected hydrocephalus during infusion tests . . . 16

Antioxidant Activity of Popular Food Products . . . 17

Application of language models in predicting protein aggregation capacity 18

Bioinformatical analysis of disease-causing proteins originating from the X and Y chromosomes . . . 19

Bioterrorism in the age of artificial intelligence . . . 20

Comparative analysis of cores of IncF bacterial plasmids from the Enterobacteriaceae family. . . . 21

Current protocols for bioinformatics analysis of scRNA-seq data and their applications 22

Enhancing COVID-19 chest X-Ray classification via artificial extension of the dataset with variational autoencoder . . . 23

Exploring Potential Side Effects of Drugs through Molecular Docking New Therapeutic Pathways in Disease Treatment . . . 24

Exploring the Impact of High Heat Styling on Hair Molecular Composition ATR-FTIR Analysis . . . 25

From Cooperation to Competition revealing stable correlations between features in the evolutionary game theory . . . 26

From Darwin to Data Popularization of Genetic Algorithms . . . 27

Genetic algorithms and their applications in bioinformatics . . . 28

GO-a-GO Functional Annotation of Genes in Chromatin Contact Pairs 29

Impact of Intracellular Amino Acid Dynamics on Chronological Life Span in Fission Yeast 30
Lactate Dehydrogenase in Humans 31
Logistic regression as a method for identification of protein signatures of non-small cell lung cancer subtypes 32
Metagenomic bioprospecting of cold-active pectinases 33
Molecular dynamics simulation (MD) Sophoraflavanones G as SGLTs inhibitors . . . 34
Proteogenomic and metabolomic characterization of human glioblastoma. 35
PyLig An In Silico Protein-Ligand Interaction Visualization Software 36
The influence of DNA-histone interactions on binding of molecules in the minor groove of DNA. 37
Understanding Transformer Models Insights and Applications 38
Using CNNs and ATAC-STARR-seq to uncover predictive opportunities in regulatory genomics 39
Vibrational spectroscopy methods in the Identification of Basidiomycota 40

Author's index **41**

PRESENTATION ABSTRACTS

Application of long-read sequencing to improve genotyping of complex pharmacogenetic regions

Maria Paleczny[1], Dżesika Hoinkis[1], Sławomir Gołda[1], Katarzyna Tomala[1], Monika Opalek[1], Klaudia Pacewicz[1], Klaudia Szklarczyk-Smolana[1], Michał Korostyński[1], Marcin Piechota[1]

[1]Intelliseq

Pharmacogenetic regions, such as the CYP2D6 gene locus, are critical for determining individual drug metabolism and response. However, the genotyping of these regions is challenging by the frequent presence of complicated structural rearrangements and in the case of CYP2D6 proximity of highly homologous pseudogene CYP2D7. Traditional short-read sequencing technologies often fall short of accurately mapping these regions due to ambiguities arising from sequence similarity. This can result in incorrect assessments of gene diplotypes in star-allele nomenclature, which subsequently leads to inaccurate metabolism level assignment and drug dosing recommendations. To address these challenges, this study explores the application of long-read sequencing technologies, including Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (Nanopore), which significantly improve sequencing read length over the complex loci. These technologies enable more precise mapping of reads, enhancing the specificity of pharmacogenetic assessments. Moreover, long-read sequencing also facilitates better phasing than short-read sequencing, which improves the identification of gene haplotypes. We developed bioinformatic workflows to optimise the processing of long-read sequencing data, focusing on alignment, error correction, and variant calling specifically tailored to handle complex genomic structures. A key component of the entire analysis is our proprietary Polygenic tool, which identifies the most probable star-allele haplotype pair. Validation of our approach was conducted using two datasets: 1) a panel of well-characterised samples from the Coriell Institute sequenced with PacBio technology, which has known pharmacogenetic genotypes, and 2) a set of samples obtained as part of the PGx Plus project, containing samples sequenced with both short-read and long-read Nanopore technology. The results prove the high accuracy of CYP2D6 genotyping with long-read sequencing compared to traditional methods, as well as the excellent quality and robustness of developed bioinformatic workflows. These findings underscore the potential of long-read sequencing in resolving complex pharmacogenetic regions, thereby facilitating the customisation of pharmacotherapy. Such advancements have the potential to contribute to the rapidly growing field of precision medicine greatly.

This work was supported by PGx Plus project No. POIR.01.02.00-00-0089/18-00.

Clinical trials. Real World Data as a tool of implementation of innovative medical technologies.

Aneta Kominek[1],

[2] *2KMM sp. z o.o.*

keywords: *clinical trials, Real World Data, data analysis*

The implementation of solutions and technologies that save lives and human health, or improve the quality of life, requires clinical trials. Clinical trials provide evidence regarding the safety and efficacy of proposed therapies. As it relates to human life and health, clinical trials are under heavy supervision of regulatory authorities. There are a number of regulations and guidelines to which all stages of a clinical trial are subjected, from the design of the trial's objectives, to the provision of reliable and validated data, the development of results and standards for their presentation. In the process of providing clinical evidence of efficacy and safety, the patient has the central place. It is the patient, the human being, who consciously takes the risk by voluntarily agreeing to conduct an experiment with his or her participation. Aware of these risks, the clinical research community is constantly looking for the best solutions that can reduce the patient's exposure to the inconvenience and risks of participating in research. Another issue is the cost of introducing new medical solutions. Thus, there is a permanent search for efficient methods to optimize the process of collecting necessary data. Dynamic technological advances, facilitating wide access to digital data, underpin the increasingly popular idea of using Real World Data, perfectly complementing the evidence gathered during tightly controlled clinical trials. We're talking about the use of information that exists in different systems, in different formats, translated to one unified format, to confirm assumed hypotheses about the effectiveness of solutions. 2KMM is an example of a Clinical Research Organization (CRO) that conducts this kind of projects, for the benefit of all clinical research stakeholders. An interesting, "trendy" and perhaps inevitable development in this area of science is the use of artificial intelligence (AI). There are already known examples of using AI to find the best promising molecules with the potential to be introduced into clinical practice. The ethical problem of human participation in experimental therapies is also being considered in the context of the use of AI. Attempts are being made to use AI to generate datasets that mimic the real population. It allows reducing the number of patients required in a trial, so a smaller group of people are exposed to the consequences of participating in an experiment. In consequence it also reduces the cost of conducting the trial. Thus, these are extremely interesting branches of science, the intensive expansion of which we are watching with great hope.

Development of WGS pipelines for computation and reporting of polygenic risk scores

Mateusz Marynowski[1], Marcin Piechota[1], Klaudia Szklarczyk-Smolana[1], Michał Korostyński[1], Klaudia Pacewicz[1], Katarzyna Tomala[1], Sławomir Gołda[1], Dżesika Hoinkis[1]

[2]*Intelliseq*

As the volume of whole genome sequencing (WGS) and low-pass sequencing data within biobanks continues to grow, using this information for effective polygenic risk score (PRS) computation presents significant challenges. Traditional GWAS studies and PRS models were predominantly developed based on microarray genotyping data. The available tools often fail to capture the comprehensive genetic variations revealed by next-generation sequencing (NGS) technologies. A primary obstacle is the transition from the GRCh37 to the GRCh38 reference genome, which also includes numerous alternate loci in clinically important regions (including HLA). Furthermore, the increasing diversity of the populations studied in recent genomic research introduces complexities in PRS calculations due to population-specific allele frequencies. More sophisticated bioinformatic tools are required to adjust for population stratification and genetic heterogeneity. To address these issues, we have designed an advanced WDL workflow capable of accurately computing and reporting PRSs. Our pipeline is anchored by a polygenic Python package, which leverages YAML-defined models to interpret and process VCF files, ensuring flexibility and adaptability to various genetic contexts. The workflow calculates more than 600 polygenic scores on four genetic backgrounds. It integrates risk scores provided by leading and trusted databases: GBK- Global Biobank Engine, Genebass- gene-based association statistics, PGS Catalog- The Polygenic Score Catalog, and GWAS Catalog- The NHGRI-EBI Catalog of human genome-wide association studies. The reporting module includes the use of Gaussian distribution graphics and percentile rankings to depict PRS outcomes, along with a color-coded system that categorizes risk levels. The visualization strategy not only makes the data more accessible to clinicians but also aids in patient communication, providing a clearer interpretation of risk scores in a clinical setting. By integrating these graphical elements, we ensure that the PRSs are not only scientifically robust but also practically useful for risk assessment and patient counseling. The workflow represents a significant advancement in the field of genomic medicine, addressing critical gaps in current methodologies and setting a new standard for the utilization of extensive genomic data in risk prediction and health management.

This work was supported by Mobigen project No RPMP.01.02.01-12-0049/18.

iFlow platform for automated interpretation and customised reporting of NGS data in genome-based medicine

Klaudia Szklarczyk-Smolana[1]

[2]*PhD, CEO of Intelliseq*

Genomic data impacts all phases of healthcare- it accelerates diagnostics, improves the effectiveness of treatment, and enables personalized prevention based on individual genetic risk. However, there are still challenges in implementing genome-based medicine into clinical practice. The analysis of already sequenced genomes is expensive and labor-intensive. In many cases there are few experts in the field, no proper IT tools, and difficulties in transferring scientific knowledge into clinical practice. Automating the entire process of analysis and interpretation is a way to revolutionize healthcare. We have developed a technology that streamlines processes in clinical laboratories, enabling precise and rapid generation of clinical reports while reducing costs. Intelliseq introduces the iFlow solution for automated analysis and reporting of human genomic data. It operates as a multi-level platform, supporting the creation of advanced analytical paths with proprietary and AI-based algorithms. The engine facilitates the generation of GeneSpect Reporter for interpretation in various areas of genome-based medicine, including somatic cancer, inherited diseases, and pharmacogenomics. Its web-based interface facilitates the user-friendly execution of analyses and report generation. We will present examples of how the Engine enables comprehensive secondary and tertiary analyses from raw data processing, through automated variant calling and annotation, to describing all clinically relevant findings. Intelliseq's Customica tool enhances reporting personalization, enabling users to tailor reports to specific requirements. Real-world applications showcase reporting capabilities in molecular diagnostics labs, Contract Research Organizations (CROs), and integrated healthcare systems. Through the iFlow solution, Intelliseq advances genomic analysis, accelerating the translation of genetic insights into actionable clinical outcomes.

This work was supported by Mobigen project No RPMP.01.02.01-12-0049/18 and PGx Plus project No. POIR.01.02.00-00-0089/18-00

Secrets of the link between neurodegeneration and bacterial amyloid migration

Oliwia Polańska[1] Monika Szefczyk[1] Małgorzata Kotulska[2]

[1]Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland
[2]Department of Bioorganic Chemistry, Faculty of Chemistry, Wroclaw University of Science and Technology, Wroclaw, Poland

Keywords: neurodegenerative diseases, gut microbiota, amyloids, interaction

Neurodegenerative diseases (NDDs) are characterized by progressive damage to nerve cells in various areas of the brain, leading to impairment of cognitive functions. They are often incurable and fatal. Alzheimer's disease (AD), Parkinson's disease (PD), and prion diseases (such as kuru) are examples of NDDs. The presence of protein aggregates with a characteristic β-cross structure (amyloids) is often a hallmark of these diseases.

Besides amyloids associated with disease entities, there are also functional amyloids found e.g. in certain bacteria. These bacteria often inhabit the human body, with their largest populations residing in the intestines, but they can also be found in the oral cavity and nose. Amyloids secreted by bacteria often serve specific functions, such as providing stability to biofilms, which enables them to resist adverse environmental conditions.

Many studies suggest a direct link between changes in gut microbiota and neurological disorders. It has been proven that amyloids associated with disease entities can interact with bacterial amyloids. One hypothesis proposes that their mutual contact is possible via the neuronal pathway. Such interactions could influence the development of neurodegenerative diseases.

This popular science lecture will serve as a platform for deepening knowledge about potential interactions between gut microbiota and the nervous system. This may contribute to a better understanding of the processes leading to neurodegenerative diseases.

POSTER ABSTRACTS

A deep dive into invariant Natural Killer T (iNKT) cells: a single-cell transcriptomics perspective

Joanna Dąbrowska[1], Emma Ingelbinck[1], Pavel Ostašov[2], Monika Holubová[2], Aleksander Jankowski[1]

[1] *University of Warsaw*
[2] *Charles University*

The invariant Natural Killer T (iNKT) cells, a unique lymphocyte subset, exhibit characteristics of both T cells and Natural Killer cells, thereby serving as an interface between innate and adaptive immunity. These cells are primarily recognized for their immunoregulatory capabilities, particularly inducing anti-tumor responses in cancers expressing CD1d. This is facilitated by their invariant T-cell receptor (TCR), which recognizes glycolipid antigens presented by the non-polymorphic major histocompatibility complex class I-like molecule.

Despite their potential, the applications of iNKT cells in cell-targeted immunotherapy of human cancer patients have been limited. To enhance the therapeutic efficacy of iNKT cells, it is crucial to comprehend the heterogeneity within this population and the functional roles of its specific subtypes.

In this context, we propose to expand the knowledge of iNKT cell biology by characterizing the known subtypes of iNKT cells. To further inspect the diversity of iNKT cells, we will employ single cell transcriptomics (scRNA-seq) data from the laboratory of Dr. Monika Holubová at Charles University. The samples were procured from four healthy peripheral blood donors, with iNKT cells isolated and cultured using two different activators, interleukin 2 or interleukin 15.

The scRNA-seq data will be utilized to cluster the assayed cells, identify markers, provide functional characteristics for individual cell subtypes, and compare how different activators influence the subtype profile of the given iNKT sample. Finally, we discuss the selection of the more promising method for future immunotherapy applications.

A new method for protein function prediction.

Szymon Koruba[1], Klaudia Skutnik[1], Nikola Pawłowska[1], Martyna Szyszka[1], Julia Merta[2], Patryk Jarnot[3]

[1] *Faculty of Environmental Engineering and Energetics, Silesian University of Technology*
[2] *Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology*
[3] *Department of Computer Network and Systems, Silesian University of Technology*

Proteins are involved in almost every biological process and regulation of living organisms. Nowadays, high-throughput next-generation sequencing leads to a vast number of known protein sequences that we are unable to analyse manually. This is challenging for scientists who use expensive and time-consuming experimental methods to discover their real functions. To speed up and reduce the cost of these methods, they predict their possible functional and structural properties, which helps them reduce the number of experimental scenarios. Therefore, developing accurate methods to automatically predict protein functions is a critical task that requires models capable of performing multi-label classification. In this research, we focus on improvements in protein function prediction models.

Our goal is to develop a new method for predicting protein functions based on protein sequences. We designed the workflow by selecting the appropriate input dataset, features of protein sequences, machine learning methods, and validation approaches. As an input dataset we use protein sequences and their known functions described by GO terms provided by the organisers of the CAFA challenge. Protein sequences hide many features that scientists are gradually discovering. These include physicochemical properties, domains and structures of proteins. Initially, we select the most common features used for this task, which are amino acid frequencies, physico-chemical properties, and residue contact information. We first compare available models with various parameters to check their impact on the quality of the results. Then, taking into account the conclusions from other solutions, we will propose our own method. Finally, we will use three different submodels with different sets of features for all three GO categories, which are: molecular function, cellular component and biological process. To validate our model, we use cross-validation with F_{max} and S_{min} scores as metrics. These metrics are used in the CAFA challenge, in which we also plan to participate.

Adopting the linear regression model to characterize the diversity profiles of TCR repertoires.

Alicja Bargieła[1], Justyna Mika[2]

[1] *Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland*
[2] *Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland.*

Introduction: Adaptive immunity relies on diverse B and T cells to mount specific responses against pathogens. T cell receptors (TCRs) ensure specificity by binding to antigens presented by infected or aberrant cells. Assessing T cell diversity involves analyzing the number of unique TCR sequences and their abundances. Hill numbers are versatile measures of diversity, which control the emphasis placed on rare TCR sequences with the use of alpha parameter, creating so called diversity profiles. Here we intend to evaluate the use of linear regression modelling of diversity profiles to describe the heterogeneity of TCR repertoires.

Methods: We used two datasets of human TCR repertoires obtained from 587 healthy donors (HD, Dean et al https://doi.org/10.1186/s13073-015-0238-z) and 21 melanoma patients (MP, Huuhtannen et al https://doi.org/10.1038/s41467-022-33720-z). Skewness of distributions was estimated with the use of discrete power law distribution. TCR diversity was defined as Hill numbers of order alpha ranging from 0 to 5 with step 0.1. Small alpha values put more emphasis on rare clones, whereas as alpha gets bigger, more focus is put on abundant TCRs. Variable transformations were employed to enhance the fitting of a linear regression model to the diversity profiles. The best transformation was chosen based on BIC criterion. Comparative study between biological conditions (Male vs Females for HD dataset and Treatment vs No Treatment for MP dataset) was performed using 95% confidence intervals for model coefficients.

Results: HD repertoires were on average more skewed than the MP repertoires. The diversity was higher for small alpha values when compared to large alpha values for both analysed datasets. Box-Cox transformation, with lambda ranging from -0.9 to 0.2 was applied before modelling the data with linear regression. Both datasets had negative relationship between the diversity and alpha parameter for all repertoires. The more skewed dataset (HD) had higher intercept coefficient values than the less skewed dataset (MP). There was no statistical difference between the biological conditions in the intercept coefficient values as well as slope coefficient for both analysed datasets.

Conclusions: Diversity profiles provide a valuable tool for observing the trajectory of changes that cannot be captured by individual diversity indices. A comparative analysis of profiles in terms of distribution skewness can show that less skewed repertoires have a relatively even count of individual TCR copies, while more skewed profiles exhibit a more diverse number of clones for each TCR. Linear regression provides a simple summary of a diversity profile, allowing for statistical comparison of different TCR repertoires

The work was supported by the SUT's grant for Support and Development of Research Potential.

Analysis of heterogeneity in hematopoietic stem cell differentiation trajectories using RNA sequencing at the single-cell level.

Anna Garbarz[1]

[1] *Jagiellonian University*

Hematopoietic stem cells (HSC) are cells that can produce all types of blood cells and maintain their own population through self-renewal. Analysing the heterogeneity of haematopoietic stem cell differentiation trajectories is crucial for understanding the origins of leukaemia, one of the most complex blood cancers. Although leukemic cancer cells' genetic heterogeneity is well defined, the precise mechanisms underlying the initiation of cancer transformation remain elusive. This process,

characterized by its long-term, complex, and multi-stage nature, poses significant challenges to the development of effective therapeutic strategies. A critical obstacle lies in pinpointing the precise stage of hematopoietic development at which cells undergo malignant transformation. Addressing this problem holds the potential to develop precise therapies that would improve the chances of survival for numerous patients.

The aim of this study was to use single-cell analysis to identify hematopoietic stem cells that could potentially undergo the initial stages of cancer transformation.

Data for single-cell sequencing analysis comes from the bone marrow of leukaemia patients.. The single-cell RNA sequencing analysis involved filtering and preprocessing the data using the Seurat library in R, followed by performing velocyto and scvelo analysis in Python. The analysis revealed the presence of several clusters of cells with unique gene expression patterns and visualised the trajectories of cell differentiation. Some clusters exhibit a tendency to favour specific cell lineages, suggesting their potential role in the process of neoplastic transformation.

The analysed diversity within the blood stem cell population, focused especially on the clusters with clear lineage bias, shows the presence of specific cell subsets that could hold a potential for the development of new therapeutic strategies in the treatment of leukaemia.

Analysis of intracranial pressure in patients with suspected hydrocephalus during infusion tests.

Monika Najdek[1], Marek Czosnyka[2], Agnieszka Kazimierska[3]

[1] *Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology*

[2] *Brain Physics Laboratory, Department of Clinical Neurosciences, University of Cambridge*

[3] *Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology*

keywords: *intracranial pressure, cerebrospinal fluid, pulse waveform, normal pressure hydrocephalus, infusion test*

The aim of this study was to analyze parameters that describe changes in intracranial pressure (ICP) related to the cardiac cycle, called the ICP pulse waveform, during controlled changes in mean ICP. The study analyzed retrospective ICP recordings collected during infusion tests performed in 29 patients with suspected hydrocephalus. Three pulsatile parameters were examined: amplitude (AMP), the slope of the ascending part of the pulse (M) and the normalized area under the curve (S), along with two compensation parameters: resistance to CSF outflow (Rcsf) and elastance coefficient.

Analysis of the relationship between analyzed parameters showed that there were significant correlations between the pulsatile parameters averaged over the three stages of the infusion test and Rcsf (Spearman correlation coefficients varied between 0.43 to 0.88, $p<0.05$), which is an indicator used commonly in the diagnosis of hydrocephalus. This means that pulse parameters investigated in this study could potentially be used to assess pressure-volume compensation and to determine whether the patient's condition will improve after shunting. Significant correlation was also detected between the time courses of mean ICP and three pulsatile parameters (Spearman correlation coefficients: 0.91 [0.82–0.96], $p<0.001$ for AMP, 0.76 [0.64–0.88], $p<0.001$ for M, and 0.3 [0.07–0.69], $p<0.004$ for S). Significant correlations between the time courses of mean ICP and M and S suggest that they could be used to continuously assess the state of the intracranial space.

Antioxidant Activity of Popular Food Products.

Maja Muszer[1], Kinga Barszcz[2], Nikola Rybarczyk[3], Zofia Dobrowolska[4], Jakub Młotkowski[4], Oliwia Polańska[4], Tomasz Walski[4]

[1] *Department of Biochemistry, Molecular Biology and Biotechnology Faculty of Chemistry, Wroclaw University of Science and Technology, Wroclaw, Poland*
[2] *Department of Optics and Photonics, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland*
[3] *Department of Experimental Physics, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland*
[4] *Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland*

Keywords: *disease, oxidative stress, reactive oxygen species, total antioxidant capacity*

As a result of natural processes occurring in living organisms and exposure to various external factors, reactive oxygen species (ROS) are formed within cells. ROS are highly reactive molecules or substances containing unpaired electrons. The accumulation of ROS in the body leads to a phenomenon known as oxidative stress.

It carries with it a range of negative consequences, potentially contributing to the development of many difficult-to-treat diseases that continue to pose significant challenges for modern medicine. These include pathologies such as cancers, neurodegenerative diseases, and autoimmune disorders. In addition to the natural processes of the body combating oxidative stress, many commonly available food products contain substances with antioxidant properties (known as antioxidants) such as vitamin C or E. Incorporating antioxidant-containing products into daily diet may contribute to reducing the risk of developing many of these diseases. Using radical scavenging assay based on UV-VIS spectroscopy and popular chemical reagents like ABTS+• and Trolox (a water-soluble analog of vitamin E), we assessed total antioxidant capacity (TAC) of various food liquids (e.g. wines, fruit juices, coffee). Our findings pinpoint potential dietary additions that could help to reduce the incidence of diseases by neutralizing ROS effectively.

Application of language models in predicting protein aggregation capacity

Paweł Tomkowski[1], Zofia Dobrowolska[2], Monika Najdek[2], Alicja Nowakowska[2]

[1]Wrocław University of Environmental Science
[2]Wrocław University of Science and Technology

Amyloids are abnormal proteins that aggregate in human and bacterial organisms to form stable fiber-like or plate-like structures. The accumulation of amyloids can lead to the formation of toxic aggregates that damage cells and can lead to neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and multiple sclerosis. Nevertheless, protein aggregation can also be a positive event e.g. in bacteria amyloids help to stabilize biofilm. As existing protein aggregation predictors are not satisfactory, especially for long proteins, developing a new tool that uses the numerical representation of proteins obtained through the ProtBERT model was addressed. In the project, we analyzed 547 protein sequences. For each protein sequence, a numerical representation derived from the language model ProtBERT was used. We attempted to differentiate proteins using unsupervised techniques including PCA and T-sne, however no clear separation was observed. Next, we performed a classification analysis with standard models such as logistic regression, SVM, random forest and decison tree. The best model achived AUC of 0.82. Despite the comprehensive data analysis and the application of various dimension reduction and classification modeling techniques, the results obtained did not stand out, in quality, compared to competing solutions.

Bioinformatical analysis of disease-causing proteins originating from the X and Y chromosomes

Julia Cieszko[1], Mikołaj Kikolski[2]

[1] *University of Wroclaw*
[2] *University of Wroclaw*

Autoimmune diseases are a group of conditions in which the body's own immune system attacks healthy cells. Although their exact cause is not fully understood, both genetic and environmental factors are suspected to play a role. HLA proteins on the surface of cells are crucial for recognizing cells as self; when there are defects in their structure, it can lead to an attack on one's own cells. There is an association between gene expression and autoimmune diseases, especially in women, suggesting the influence of sex hormones on the immune response. Genetic and bioinformatic studies reveal altered gene expression related to the immune system in affected individuals. Analyzing gene expression data in patients with rheumatoid arthritis (RA) and systemic lupus erythematosus (SLE) has shown differences between sexes both in susceptibility to diseases and in gene expression associated with them. Genes located on the X and Y chromosomes have been found to be involved in the pathogenesis of these diseases. Studies suggest that signaling pathways related to the immune response and sex hormones may be key to the development and course of these diseases. Identification of genes and molecular pathways may lead to a better understanding of the pathogenic mechanisms of these diseases and the identification of potential therapeutic targets.

Bioterrorism in the age of artificial intelligence

Jakub Bogacz[1], Dominika Porzybót[1],

[1] *University of Wroclaw*

Rapid advances in artificial intelligence (AI) have raised growing concerns among experts, policymakers and world leaders about the potential for increasingly sophisticated AI systems to pose catastrophic risks. One threat that particularly focuses attention is the potential of these tools to support acts of bioterrorism. The level of risk that AI-bio capabilities may pose for biosecurity, and which tools pose the greatest risk, are the subject of discussion within the life sciences and AI expert communities. Our review attempts to summarize the current state of this discussion, describes trends in preventing the use of AI-bio models by malicious actors, offer practical suggestions for mitigating these dangers, foster a comprehensive understanding of these risks and inspire proactive efforts. We pay particular attention to the potential use of large language models as an assistive tool in the biohazard creation, ideation, magnification and release process.

Comparative analysis of cores of IncF bacterial plasmids from the Enterobacteriaceae family.

Anastazja Tasinkiewicz[1], Agnieszka Prudło[1], Stanisław Gołębiewski[1]

[1] *University of Warsaw*

Genomic data are key a source of information in biology. The detailed analysis of these data allows us to draw conclusions about the characteristic features of the examined organism and its potential abilities. In turn, comparative analysis additionally enables understanding of the evolutionary processes that occurred among the examined sample and their biological diversity observed as a result. In bacteria, one of the main sources of high variability are mobile genetic elements, which include, among others, plasmids. In our work, we focus on the comparative analysis of cores of selected plasmids from the Enterobacteriaceae family. We take a closer look at their ori sites and the determinants of incompatibility - replication and partition systems. On this basis, we draw conclusions about the details of the mechanisms necessary for the survival of the studied mobilomes in nature. Moreover, we consider their possible evolutionary connections. We also suggest the next steps of such comparative analysis focused on plasmid cores, which may serve as an introduction to their potential automation. The aim of our work is to obtain in-depth information about selected plasmids based on the analysis of their cores. As well as proposing a sketch plan for creating a bioinformatics algorithm oriented towards plasmid cores.

Current protocols for bioinformatics analysis of scRNA-seq data and their applications

Maria Koziarz[1], Julia Kwiecińska[1], Adrian Kania[1]

[1] *Jagiellonian University*

Single-cell RNA sequencing (scRNA-seq) technology has become the state-of-the-art approach to discovering the diversity and complexity of RNA transcripts in analyzed cells. Among other applications, it can be used to discover new cell types and their functions in highly organized tissues, organs or whole organisms. We present the key elements and current procedures for analyzing scRNA-seq data with a special focus on the bioinformatics part. We outline one of the most promising applications of scRNA-seq technology, which is the construction of a better and high-resolution catalog of cells in all living organisms, commonly known as an atlas. It provides an important source for a deeper understanding of cellular function, which will probably result in improvement of treating and easier diagnosis of diseases.

Enhancing COVID-19 chest X-Ray classification via artificial extension of the dataset with variational autoencoder

Urszula Cojg[1], Aleksandra Suwalska[2],

[1]Faculty of Biomedical Engineering, Silesian University of Technology, Franklina Roosevelta 40, 41-800 Zabrze, Poland
[2]Department of Data Science and Engineering, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland

Keywords: *X-Ray, Neural Network, Variational Autoencoder, COVID-19, Deep learning*

Introduction: The global COVID-19 pandemic, caused by the SARS-CoV-2 virus, has highlighted the importance of chest X-ray images (CXR). The use of X-ray images to detect and diagnose COVID-19 has become a vital tool in the battle against the worldwide crisis. However, when constructing neural network models to classify patients with different conditions, the challenge of imbalanced datasets emerged. Imbalanced datasets often lead to incorrect predictions, particularly for minority cases. The aim of this work was to compare the effects of balancing class sizes in medical image classification problems using neural networks.

Materials and methods: The study used the BIMCV COVID-19+ dataset from the Valencian Region Medical ImageBank, which contains chest X-ray images of patients infected with the virus and their radiological findings and locations. To prepare the data, segmentation of lungs was performed, and the high-resolution images were downsampled to a resolution of 256x256 pixels. The dataset was split into a training set (70%) and a test set (30%) with a consistent ratio of categories. A variational autoencoder was used to reduce dimensionality and generate low-dimensional representations for each image. Synthetic data augmentation methods such as Borderline-SMOTE were used to balance class sizes. In the final classification, the data was passed through a fully connected neural network. The trained model was evaluated using several metrics, including balanced accuracy, sensitivity and specificity.

Results: The neural network constructed based on the variational autoencoder model consists of three convolutional layers, each followed by a ReLU activation function. For classification, a fully connected neural network with two fully connected layers was employed. The softmax function was applied to transform the results into probabilities for each class. Balanced accuracy after synthetic data augmentation reached 75.05%, indicating that the network accurately predicted most COVID-19 patients, compared to 66.89% with imbalanced data. Prior to data augmentation, specificity was 53.99%, which improved to 71.39% after augmentation. Sensitivity results were close both before (79.80%) and after (78.71%) augmentation.

Conclusion: In summary, the proposed data augmentation method using the Variational Autoencoder for COVID-19 patient classification tasks stands as a noteworthy and valuable addition to the field of medical imaging and disease diagnosis. The experiments conducted in the study demonstrated the usefulness of the augmentation approach for challenging problems with a limited number of cases. Such an improvement could benefit healthcare and enable faster responses to public health threats, including global pandemics.

Exploring Potential Side Effects of Drugs through Molecular Docking: New Therapeutic Pathways in Disease Treatment

Katarzyna Nykiel[1], Urszula Jaśnikowska[1], Patrycja Gonciarz[1]

[1] *University of Wroclaw*

In today's rapidly evolving world of medicine, increasing attention is being paid to the search for new therapies for a variety of diseases. In this context, the use of molecular docking techniques becomes incredibly attractive due to its ability to identify potential side effects of drugs, which may lead to the discovery of new therapeutic pathways. The aim of our work is to investigate the potential of utilizing molecular docking techniques in identifying these potential side ef- fects that may have a beneficial impact on the treatment of other diseases.

Our work is based on analyzing the operation of advanced computer tools that al- low for the analysis of drug structures. Through docking simulations, we strive to identify new drug applications by analyzing their interactions with target proteins. Many existing drugs may have potential in treating diseases different from their original intended use. Molecular docking methods support research on anticancer drugs and are helpful in creating drugs against resistant bacterial infections. Discovering potential side effects of drugs through molecular docking opens new perspectives in disease treatment and can significantly accelerate the development of more effective therapies. However, it is important to note that further research and experiments are necessary to confirm our observations. Po- tential outcomes of this research could include the identification of novel drug candidates for repurposing in the treatment of various diseases, thereby ex- panding the therapeutic options available to patients. Additionally, the insights gained from molecular docking may lead to the development of more targeted and efficacious therapies, ultimately advancing precision medicine approaches in healthcare.

Our work aims to draw attention to the potential inherent in the application of molecular docking techniques to search for new therapies that may have a significant impact on improving the health and quality of life of patients.

Exploring the Impact of High Heat Styling on Hair Molecular Composition: ATR-FTIR Analysis

Nikola Rybarczyk[1], Kinga Barszcz[2], Oliwia Polańska[3], Marlena Gąsior-Głogowska[3],

[1]Department of Experimental Physics, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland
[2] Department of Optics and Photonics, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland
[3] Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland

Keywords: *spectroscopy, hair damage, high temperature*

Hair styling using curling irons or straighteners has surged in popularity. These devices, renowned for their ability to manipulate hair texture, rely on elevated temperatures to achieve desired results. However, the impact of such high temperatures on the molecular structure of hair remains a subject of considerable interest and concern.

This study meticulously examines the intricate relationship between temperature and the molecular composition of hair. By employing the advanced Attenuated Total Reflectance Fourier Transform Infrared (ATR-FTIR) technique, known for its remarkable sensitivity and rapid analysis capabilities, the subtle chemical alterations occurring within hair fibers when subjected to heat has been explored. Through the acquisition of spectra, which represent the vibrational modes of molecular bonds, the study discerns significant changes within water content, lipids, and proteins present in the hair strands.

The changes in position and bandwidth of band to thermal stress enable precise identification of alterations in various components, shedding light on the mechanisms underlying both hair damage and resilience during styling practices and temperature selection to avoid changes in hair composition. Additionally, these findings contribute to the development of informed hair care strategies, emphasizing the importance of temperature control and protective measures to mitigate potential damage. This knowledge enables conscious decisions, balancing desired aesthetics with the preservation of hair health and integrity.

From Cooperation to Competition: revealing stable correlations between features in the evolutionary game theory

Karol Chądzyński[1], Jakub Giezgała[1], Michał Pszenicyn[1], Karina Kołodziejczyk[1]

[1] University of Warsaw

Evolutionary Game Theory (EGT) is an interdisciplinary field of research that combines game theory with the principles of evolutionary biology. Using mathematical methods, we can try to describe how biological species adapt depending on interactions between them and imposed environmental conditions. By representing the concept of evolution as nondeterministic mutations, we can analyse the behavioural strategies observed as a result of the adaptation process. Each player (individual), with a set of immutable traits, always follows the optimal sequence of decisions in order to maximise the probability of winning the highest possible reward. Individuals, whose strategies prove to be more rewarding than others', will have the chance to reproduce, therefore increasing the number of individuals of their kind. By applying mathematical models, the EGT analyses the dynamics of strategy propagation. Our work looks not only at the values of the traits around which individuals cluster, or how changes in such clusters occur, but also at finding stable relationships between independent traits. In the simulation, individuals are randomly paired up, with three interaction options available: cooperation, aggression, or escape. Additionally, individuals may recognize the values of certain traits in their opponent and base their decisions on these. The simulation aims to demonstrate that in EGT, stable correlations between traits may exist, which can have a significant impact on evolutionary advantage. Using statistical methods and probability estimation, we wanted to predict the existence of highlighted trait values and find correlations between them among a group of individuals. Finally, we validated all our predictions by running the simulation, which allowed us to discover the actual correlations of traits in individuals. Afterwards we attempted to categorise them in order to determine the impact and quantity of individual relationships.

From Darwin to Data: Popularization of Genetic Algorithms

Olga Wieromiejczyk[1], Anna Krzywiecka[1], dr hab. Guillem Ylla[1]

[1] Jagiellonian University in Kraków, Faculty of Biochemistry, Biophysics and Biotechnology

Genetic algorithms (GAs) stand out due to their robustness in handling optimization tasks. Our poster introduces the fundamental concepts of GAs and illustrates their application within bioinformatics. In this poster, we provide a comprehensive explanation of how GAs operate. Furthermore, we showcase the genetic algorithm's ability to optimize complex biological tasks applying it to the optimization of primer design for polymerase chain reaction (PCR). Inspired by a twenty-year-old publication [1] we re-implemented a GA code to optimize the qPCR primer design and included several improvements to the original algorithm. Among others, we enhanced its computational efficiency by allowing multi-threading, and added new functionalities such as individual melting temperature range definition and adjustable length range. This GA application is an example of how GAs effectively navigate the vast solution space to identify primers that meet critical specifications such as melting temperature, primer length, GC content, and specificity. The GA application presented in this poster shows the potential of GA for solving optimization processes that can be applied to various fields of biological research.

[1] Jain-Shing Wu, Chungnan Lee, Chien-Chang Wu, Yow-Ling Shiue, Primer design using genetic algorithm, Bioinformatics, Volume 20, Issue 11, July 2004, Pages 1710–1717, https://doi.org/10.1093/bioinformatics/bth147

Genetic algorithms and their applications in bioinformatics

Aleksander Janowiak[1], Kupidura[1], Zuzanna Kiczak[1]

[1] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Stefana Banacha 2, 02-097, Warsaw, Poland

The ability for efficient data analysis is currently a major bottleneck in biological research. Machine learning algorithms have proven particularly effective in addressing this issue. Genetic algorithms, a notable class of machine learning methods, provide a biology-inspired approach based around the principle of natural selection. Their usual implementations involve the creation of an initial population of solutions to a given problem, which is then randomly mutated with mutants being evaluated according to a predefined fitness function all in a process mimicking biological evolution. Their ability to find solutions to problems with little prior knowledge, combined with the possibility of including existing knowledge through designing appropriate initial population and fitness function has made them particularly effective tools in the rapidly developing field of bioinformatics. Genetic algorithms have already been applied to several problems in bioinformatics, and shown to match and even outperform prior methods. Such problems include predicting essential genes for an organism, predicting promoter regions in genes, as well as modelling disease risk and discovering new drugs. In this poster we aim to explain the underlying principles of genetic algorithms and present an overview of their applications in genomic, transcriptomic and proteomic research.

GO-a-GO: Functional Annotation of Genes in Chromatin Contact Pairs

Daryna Yakymenko[1,2], Aleksandr Jankowski[3], Teresa Szczepińska[2]

[1] *Faculty of Biochemistry, Biophysics and Biotechnology, Jagiellonian University of Cracow, 30-387 Cracow, Poland*
[2] *CEZAMAT, Warsaw University of Technology, 02-822 Warsaw, Poland*
[3] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, 02-097 Warsaw, Poland*

Analysis of overrepresentation of Gene Ontology (GO) terms in a set of differentially expressed genes is a standard approach to investigate functional associations of gene expression changes. We developed GO-a-GO tool to be able to analyze functional terms that are overrepresented in a set of gene pairs. This provides the opportunity to see which functions are associated with gene pairs from a selected group of chromatin contacts, such as cell type specific contacts.

We used GO-a-GO to analyze GO terms enriched in top 5% of differential contacts (at distances up to 5 Mb) in mouse embryonic stem cells and neocortical neurons measured using Hi-C genome architecture mapping (Bonev et al., 2017). The analysis revealed several enriched functional terms associated with both genes in contact pairs. In summary, we developed a promising tool to study the function of chromatin three-dimensional structural contacts, which can be used independently on the type of experimental method used to measure the contacts and on the nature of contacts (e.g. loops or spatial gene clusters).

Impact of Intracellular Amino Acid Dynamics on Chronological Life Span in Fission Yeast

Karina Kołodziejczyk[1], Anastazja Tasinkiewicz[1], Anna Korczyńska[1]

[1] *University of Warsaw*

Understanding the connection between aging and changes in amino acid levels inside cells is crucial for unraveling how lifespan is regulated. While it’s known that adjusting amino acid levels can impact how long cells and organisms live, we still don’t fully grasp how these levels change as cells age. In the study presented on our poster, titled "Amino Acids Whose Intracellular Levels Change Most During Aging Alter Chronological Life Span of Fission Yeast," researchers examined the levels of free amino acids as fission yeast cells aged.

The study compared normal cells with long-living mutant cells lacking the Pka1 protein. The results showed noticeable changes in amino acid levels as cells aged. Total amino acid levels decreased as wild-type cells aged, although less so in the mutant cells. Specifically, two amino acids, glutamine and aspartate, showed significant changes: glutamine decreased notably in wild-type cells, while aspartate increased, especially in the mutant cells.

Additionally, the study revealed how these changes affected the cells’ lifespan. Adding glutamine extended the lifespan of normal cells but not the mutant ones, while adding aspartate shortened the lifespan of mutant cells without affecting normal cells.

Overall, these findings highlight specific amino acids as potential markers of aging, with their levels influencing how long cells live. This poster sheds light on the complex relationship between amino acids and aging, providing insights into potential approaches for promoting healthy aging and longevity.

Lactate Dehydrogenase in Humans, Plants and Bacteria: Structure, Function and Clinical Implications

Amelia Kurasińska[1]

[1] *Jagiellonian University*

Lactate dehydrogenase (LDH) is a ubiquitous enzyme found in organisms across kingdoms, playing a crucial role in metabolic processes.

In humans, LDH is a tetrameric enzyme involved in anaerobic metabolism, catalyzing the interconversion of lactate and pyruvate with the concomitant conversion of NAD+ to NADH. Human LDH comes in five isoforms formed from two common subunits and present in different tissues. Dysregulation of LDH activity is implicated in numerous pathological conditions, including cancer. Recent advances in structural biology and pharmacology unveiled some of the molecular mechanisms underlying LDH function, setting the stage for possible therapeutic advancements in LDH-related diseases.

In plants, LDH participates in anaerobic metabolism, particularly converting pyruvate to lactate during fermentation. It is present in all green land plants and contributes to plant stress responses, highlighting its multifaceted roles in plant growth, development, and adaptation to changing environmental conditions. Despite all this, much is yet to learn about the specific roles of LDH in different plants and the details concerning its activity and structure.

In bacteria, LDH enzymes display remarkable structural diversity and catalytic versatility, reflecting their adaptation to diverse ecological niches and metabolic lifestyles. Bacterial LDHs are involved in fermentation, respiration, and redox homeostasis, playing critical roles in energy metabolism, virulence, and antibiotic resistance. Elucidating the structure-function relationships of bacterial LDHs holds promise for the development of novel antimicrobial strategies targeting essential metabolic pathways.

Overall, there is clear evolutionary conservation as well as functional versatility of LDH across organisms, emphasizing its fundamental importance in cellular metabolism, adaptation, and disease pathogenesis. Understanding the structure and molecular mechanisms governing LDH function in humans, plants, and bacteria offers insights into the significance of this enzyme, with implications for biotechnology and medicine. Bioinformatics is one of the viable ways to help shed light on the yet unknown properties of this omnipresent enzyme.

Logistic regression as a method for identification of protein signatures of non-small cell lung cancer subtypes

Patrycja Podleśny[1] Joanna Tobiasz[2,3]

[1]Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland
[2]Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland
[3]Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland

Keywords: *lung cancer; TCGA; proteomics; logistic regression; feature selection*

Introduction: Approximately 70% of lung cancer cases are classified as lung squamous cell carcinoma (LUSC) and lung adenocarcinoma (LUAD), which makes them two major subtypes. Since many biological functions are mainly executed by proteins, the proteome is essential for detecting biomarkers that can help with patient diagnosis and subsequent care. The transcriptomic and genomic analyses do not always provide an exact reference for protein levels and activities. Therefore, direct examination of the functional proteome can provide information that complements and extends the genomic, epigenomic, and transcriptomic analysis. This study aims to identify proteins that could differentiate LUAD and LUSC subtypes.

Methods: Identifying potential biomarkers of non-small cell lung cancer (NSCLC) subtypes involved collecting normalized protein levels from The Cancer Genome Atlas (TCGA) project, measured through Reverse Phase Protein Arrays (RPPA). The dataset included records for 216 proteins and 682 samples (LUSC n=322, LUAD n=360), 30% of which were selected as the validation set. Preliminary feature filtration was conducted using Glass rank biserial correlation effect size (ES), Mann-Whitney U test, and Gaussian Mixture Model (GMM) decomposition. Protein biomarkers determined by $|ES| > 0.1$ were used to build logistic regression models within the multiple random cross-validation with 100 repetitions. The training and test sets comprised 70% and 30% of the remaining data, respectively. In each repetition, backward feature selection with AIC was used to create the model. Quality measures were calculated for the cross-validation outcome models, and feature ranking was conducted using an accuracy measure and the order of adding features to each model.

Results: Two proteins (SMAD4 and ACC_pS79) showed particularly high values in feature ranking compared to other proteins. SMAD4 mutation can be used to identify NSCLC patients with poor survival, and the gene itself represents a potential therapeutic target in non-small cell lung cancer. Inhibitors of the protein ACC_pS79 enzyme prevent tumor growth, confirming the importance of ACC in lung cancer. However, quality measure values indicate those proteins alone are insufficient to distinguish between the subtypes.

Conclusion: The results suggest that protein expression is an important factor modulating the behavior of lung cancer cells, varying between the subtypes. These expression changes offer opportunities to target therapeutic interventions and gain deeper insights into lung cancer's mechanisms. Future research should delve into characterizing NSCLC proteomic profiles more comprehensively. Moreover, integrating proteomic data with genomic and transcriptomic analyses is essential for a more comprehensive understanding of NSCLC.

Metagenomic bioprospecting of cold-active pectinases

Michał Stanowski[1]

[1] *University of Warsaw*

Cold-active enzymes have gained considerable attention in recent years due to their potential applications in various industries. Among these enzymes, pectinases play a crucial role, particularly in the food and paper industries. The efficiency of cold-active pectinases at lower temperatures makes them desirable for processes once conducted at higher temperatures, now feasible at lower ones, offering advantages such as energy savings. Here, the algorithm for extracting cold-active pectinase genes from metagenomic data is presented, along with an experimental confirmation of their enzymatic activity in a wet laboratory. The study commenced with the extraction of raw metagenomic data from environmental samples collected from Svalbard. Subsequently, predicted genes were filtered to identify potential pectinases based on length and the requirement for full-length sequences with start and stop codons. These sequences underwent further screening using Hidden Markov Models (HMM) protein domain profiles. Finally, representative sequences from clusters obtained through MMseqs clustering were subjected to protein structure prediction using AlphaFold. After identifying potential pectinase genes, the study proceeded to the laboratory confirmation phase. This involved de novo synthesis of the selected genes in the pET expression vector. Transformed expression strains were assayed with Congo red dye. Subsequently, the enzymes underwent purification, and their enzymatic kinetics was characterized. While final results are pending, preliminary findings suggest the efficacy of the algorithm and overall approach in identifying potential cold-active pectinase genes from metagenomic data, paving the way for future advancements in industrial applications.

Molecular dynamics simulation (MD) Sophoraflavanones G as SGLTs inhibitors

Nadhiri Khalidi[1] Adam Gruda[1] Patrycja Raczkowska[1]

[1] *University of Wroclaw*

Sophoraflavanone G is a flavonoid extract from Sophora flavescens a plant known for its promising medicinal properties. Sophoraflavanone G is an im- portant sodium-glucose cotransporter (SGLT) inhibitor with pharmacological significance. Sophoraflavanone G has demonstrated several biological activities, including anti-inflammatory, antioxidant, and antiproliferative effects. The compound has also shown potential in alleviating conditions such as Parkinson's disease, pulmonary fibrosis, neuroinflammation, and anti-cancer activities in modulating immune responses. Additionally, Sophoraflavanone G has demonstrated po- tent antibacterial activity and multidrug resistance behaviors by inhibiting cell wall synthesis, inducing hydrolysis, and preventing bacteria from synthesizing biofilms. Sophoraflavanone G exhibits inhibitory activities against both SGLT1 and SGLT2 and therefore studying this compound as an SGLT inhibitor is essen- tial due to the increasing prevalence of type 2 diabetes and the need for novel therapeutic approaches to manage this condition effectively. Due to their abil- ity to reabsorb glucose in the kidney SGLTs have become promising targets for diabetes treatment. By inhibiting SGLTs, we can ensure reduced glucose reabsorption, lowering blood glucose levels, and improving glycemic control in patients with diabetes.

Investigation studies on Sophoraflavanones G as an SGLT inhibitor through molecular dynamics (MD) simulations could provide a researcher the ability to explore its potential as a novel antidiabetic agent, by investigating its molecular interactions, binding affinity, and dynamics within the SGLT binding pocket.

Overall, this computational approach to studies on Sophoraflavanones G of- fers the possibility for expanding the range of therapies and advancing precision medicine strategies in diabetes management.

Proteogenomic and metabolomic characterization of human glioblastoma.

Agnieszka Michalak[1]

[1] *University of Warsaw*

The genomic approach to cancer treatment is finding the mutated driver genes. In some instances, it works sufficiently well. However, there are many cases where the developed drug is a good match for a driver gene mutation however, the efficacy is low. The reason for that is the complexity of cancer biology. Genomics does not provide information on the regulatory steps that lead from the gene to the functional protein. This gap might be bridged by integrating genomic, transcriptomic, and proteomic data, a strategy called proteogenomics. Proteogenomics gives an enormous range of possibilities for discovering new biological insights, such as: classifying molecular subtypes, identifying pathways associated with a disease, effects of genomic events in proteomic patterns (e.g. copy-number alterations).

Glioblastoma is the most aggressive brain tumor. Integrating various multi-omic data helps understand the mechanism behind cancerogenesis, interactions in the tumor microenvironment, and especially important for patients, precise identification of tumor variants for effective and personalized therapy.

PyLig: An In Silico Protein-Ligand Interaction Visualization Software

Damian Pietron[1] Somanath Shyamsundar[2]

[1] University of Wroclaw
[2] Dept of Bioinformatics, Pondicherry University

PyLig is a collaborative open-source project aimed at developing a user-friendly and cross-platform software tool for the visualization of protein-ligand interactions. The motivation behind this project stems from the observed limitations of existing software solutions in the market, which often lack comprehensive coverage of all possible interactions and exhibit inconsistent outcomes due to their respective algorithmic methodologies. Furthermore, there is a recognized shortage of efficient and user-friendly open-source tools tailored specifically for this purpose.

PyLig employs a variety of pre-existing Python libraries, including Py3Dmol and RDkit, to facilitate visualization. The backend development of PyLig is predicated on the scientific principles governing each bond, ensuring that each bond prediction algorithm is meticulously crafted in accordance with these criteria. Specifically, each algorithm is developed individually and uniquely to satisfy all scientific requirements. The front-end of the software is constructed using PYQT5 as the foundation, providing a user-friendly interface for researchers and scientists alike.

The software is designed to accept various structural file formats (e.g., .PDB, .XML, .mmCIF) as input and provide a comprehensive visualization of the potential interactions between the ligand and the protein, including hydrogen bonds, hydrophobic interactions, van der Waals interactions, ionic interactions, disulfide bridges, and pi-pi interactions.

The project is currently divided into four phases, with the first phase focusing on the development of a basic graphical user interface (GUI) and the prediction and visualization of hydrogen bond interactions. Subsequent phases will introduce advanced GUI features, expanded interaction prediction capabilities, and the integration of machine learning modules for additional functionalities, such as active site prediction and potential ligand prediction.

The initial development phase has been successfully completed, showcasing the software's ability to display three-dimensional, rotatable visualizations of hydrogen bond interactions. The GUI has also been enhanced to support multiple tabs and file management, as well as data exportation options in various formats, such as Excel and JSON. These advancements demonstrate the project's commitment to providing users with a versatile and customizable interface to optimize their interaction with the software.

The PyLig project, undertaken by two student developers, Somanath Shyamsundar and Damian Pietro ń, seeks to address these shortcomings by consolidating disparate solutions into a singular platform. By continuing to expand the capabilities of PyLig and fostering a collaborative spirit, the project aims to contribute meaningfully to the advancement of science, pushing the boundaries of protein-ligand interaction visualization and inspiring others to engage in similar endeavors that benefit the scientific community and humanity as a whole.

The influence of DNA-histone interactions on binding of molecules in the minor groove of DNA.

Sylwia Bożek[1]

[1]*Jagiellonian University in Kraków, Faculty of Biochemistry, Biophysics and Biotechnology, Cell Biophysics Department*

DNA interacts with a histone octamer to form the structure known as a nucleosome. Electrostatic interactions dominate and occur between negatively charged DNA and positively charged amino acid residues, mainly arginine and lysine. The deoxyribonucleic acid molecule interacts not only with histones but also with other proteins such as transcription factors. Furthermore, it contains a minor and major groove. The smaller groove is a binding site for small molecules including drugs and fluorescent dyes such as Hoechst 33342 and DAPI.

The study aimed to determine how much of the DNA surface area is exposed and how the structure of the DNA minor groove changes after removing proteins. This may affect the interactions of this area with small molecules.

The following methods were used: SASA (solvent-accessible surface area) analysis and measurements of the width of the DNA minor groove. The first analysis involved calculating the solvent-accessible surface area of DNA present in the nucleosome and DNA without a histone core. The fluorescent dyes Hoechst 33342 and DAPI were assumed as solvent molecules. The second analysis involved calculating the width of the DNA minor groove in the PyMOL program. Measurements were made between three different pairs of atoms. In laboratory experiments, dyes binding in a minor groove of DNA were used to determine how the amount of bound molecules changes after partial and complete degradation of proteins in HeLa cells. During protein degradation, the intensity of fluorescence from the dye bound to the DNA in the minor groove changed with time. First, the intensity increases and then decreases dramatically. Bioinformatics studies have shown that after removing the proteins, an additional DNA surface that previously interacted with the proteins is exposed. The width of the minor groove decreased following the removal of proteins.

In conclusion partial degradation of histones exposes the surface of DNA to small molecules, e.g. drugs, which is confirmed by SASA analysis and an increase in fluorescence intensity in experiments. Further degradation disrupts the stability of the nucleosome structure, as electrostatic forces cause DNA to repel each other. As a consequence, there is a change in the shape and width of the minor groove. These changes are significant because the molecules dissociate from the DNA. The current work is focused on assessing the impact of changes in the minor groove width on the quantum yield of fluorescence of molecules exhibiting affinity to DNA.

Understanding Transformer Models: Insights and Applications

Paweł Rakoczy[1], Ignacy Berent[1]

[1]Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology

Introduction: Recently, there has been increasing interest in the basic mechanisms of neural networks. However, with the growing pace of development of large language models (LLMs), a more advanced knowledge becomes necessary for a full understanding of their functioning. For many years, transformers have been the dominant architecture among LLMs, making understanding their mechanisms a key aspect of expanding our knowledge of advanced structures. It is worth noting that despite the recent emergence of models with diverse and more complicated architectures, such as cloude or mamba, transformers still remain a significant subject of research and development in this field. Currently, transformers are applied in various domains, including natural language processing, image analysis, speech synthesis, recommendation systems, and many others, highlighting their importance and versatility.

Methodology: Our work aims to present in an intuitive way the internal structure and mechanisms of transformers. Through this exploration, our goal is to enable a deep understanding of these structures, which is a crucial starting point for better understanding advanced artificial intelligence models and for creating innovative solutions based on this technology. Explaining the difference between standard neural networks and LLMs will also help better understand the possibilities and limitations of currently developing technologies in this area.

Visualization of the operation of transformers can play a key role in understanding them. In our approach, we use a spreadsheet in Excel to illustrate the operation process of the ChatGPT-2 model. This tool allows for a clear presentation of the steps taken by the model during data processing.

As part of our analysis, we will also compare the differences between different types of transformer architectures and their advantages in the context of specific applications. Additionally, we will compare transformer-based models with traditional neural network structures and recurrent neural networks to identify their strengths and weaknesses. This multidimensional analysis will allow for a fuller understanding of the role of transformers in the field of artificial intelligence. Thorough understanding of LLM's functioning can be also a starting point for deciphering "black boxes" using eXplainable AI (XAI).

Potential Results: We expect that this poster could help build deeper intuition about mechanism of transformers what may contribute to better application of models based on that technology among participants of our presentation.

Using CNNs and ATAC-STARR-seq to uncover predictive opportunities in regulatory genomics

Anna Szymik[1] Jakub Giezgała[1] Magdalena Machnicka[1] Torgeir Rhoden Hvidsten[2] Bartosz Wilczyński[1]

[1] *University of Warsaw*
[2] *Norwegian University of Life Sciences*

The field of regulatory genomics has been revolutionised by the advent of high-throughput sequencing technologies, generating a vast amount of data that provides a readout, albeit indirect, from the activity of the complex regulatory networks of the cell. One such technology is ATAC-STARR-seq, an approach that combines Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq) with Self-Transcribing Active Regulatory Region sequencing (STARR-seq). This combination gives us access to substantial amounts of measurements, however its potential for predicting regulatory activity of genomic regions is still largely uncertain.

Despite the wealth of data and our general understanding of the fundamental principles governing the activity of enhancers and promoters, specific activity function for any given DNA sequence still remains unknown. Here, we present an exploration of a convolutional neural network's (CNN) capabilities in terms of learning and predictive abilities for these sequences.

CNNs, with their ability to extract patterns from complex data, offer several advantages in the context of DNA sequence analysis. The network we used was developed by Marlena Osipowicz as a part of her master's thesis [Osipowicz, 2021]. Its architecture is based on a Basset model, designed to predict cell-specific functional activity of sequences [Kelley et al., 2016]. The data comes from two publications presenting the results of ATAC-STARR-seq experiments conducted on the same GM12878 lymphoblastoid cell line in 2018 [Wang et al., 2018] and in 2022 [Hansen and Hodges, 2022].

Our initial analysis involved a four-class classification of the DNA sequences – predicting both their regulatory activity (active/inactive) and genomic context (promoter/non-promoter). However, the resolution of the latest data at the level of tens of base pairs posed a challenge for making such detailed predictions. We then shifted our focus to a less complex classification task, predicting only the positively active, negatively active (silencing), and inactive sequences. This approach proved to be much more promising, yielding significantly improved results.

In summary, while the task of accurately predicting regulatory sequences remains a challenge, this work demonstrates the potential of machine learning methods, particularly CNNs, in advancing our understanding of regulatory genomics.

References:

Hansen, T. J., and Hodges, E. (2022). ATAC-STARR-seq reveals transcription factor–bound activators and silencers within chromatin-accessible regions of the human genome. *Genome Research, 32(8)*, 1529–1541.

Kelley, D. R., Snoek, J., and Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome research, 26(7)*, 990–999.

Osipowicz, M. (2021). Identification of chromatin regions active in human brain using neural networks. *Master's thesis, Uniwersytet Warszawski Wydział Matematyki, Informatyki i Mechaniki, Warszawa.*

Wang, X., He, L., Goggin, S. M., Saadat, A., Wang, L., Sinnott-Armstrong, N., Claussnitzer, M., and Kellis, M. (2018). High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nature Communications, 9(1).*

Vibrational spectroscopy methods in the Identification of Basidiomycota

Maria Zdankiewicz[1] Jakub Młotkowski[1] Marlena Gąsior-Głogowska[1]

[1]*Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Science and Technology, Wroclaw, Poland*

Basidiomycota includes fungi species which are common dietary staples in various cultures around the world. However, many species, when ingested, can cause serious health damage and even lead to death. This highlights the need to develop an effective identification method to minimize the risks associated with self-harvesting mushrooms for consumption.

The purpose of this study is to evaluate the feasibility of identifying Basidiomycota using two different oscillatory spectroscopy techniques: ATR-FTIR (Attenuated Total Reflectance Fourier Transform Infrared) and Raman spectroscopy. A comprehensive analysis of the spectra was carried out, and the data obtained was compared with literature standards to evaluate the capabilities of these techniques in recognizing the characteristic features of Basidiomycota to distinguish edible from poisonous species. Obtained spectra of 17 species (panther cap, penny bun, bovine bolete, and others) were analyzed in range from 1800 cm-1 to 900 cm-1 and their second and fourth derivatives were determined. The Chemometric analysis of spectra and their derivatives was preformed using Principal Component Analysis. The visualization of its results showed that method based on ATR-FTIR spectra has promise in differentiation between mushroom species, even if they belong to the same family. The results point towards the development of faster and more effective methods for distinguishing between edible and poisonous mushrooms, which could help reduce poisonings and improve food safety.

Author’s index

Bargieła, Alicja, 14
Barszcz, Kinga, 17, 25
Berent, Ignacy, 38
Bogacz, Jakub, 20
Bożek, Sylwia, 37

Chądzyński, Karol, 26
Cieszko, Julia, 19
Cojg, Urszula, 23
Czosnyka, Marek, 16

Dobrowolska, Zofia, 17, 18
Dąbrowska, Joanna, 12

Garbarz, Anna, 15
Giezgała, Jakub, 26, 39
Gonciarz, Patrycja, 24
Gołębiewski, Stanisław, 21
Gruda, Adam, 34
Gąsior-Głogowska, Marlena, 25, 40

Holubová, Monika, 12
Hvidsten, Torgeir Rhoden, 39

Ingelbinck, Emma, 12

Jankowski, Aleksander, 12
Jankowski, Aleksandr, 29
Janowiak, Aleksander, 28
Jarnot, Patryk, 13
Jaśnikowska, Urszula, 24
Joanna Tobiasz, 32

Kania, Adrian, 22
Kazimierska, Agnieszka, 16
Khalidi, Nadhiri, 34
Kiczak, Zuzanna, 28
Kikolski, Mikołaj, 19
Korczyńska, Anna, 30
Koruba, Szymon, 13
Kotulska, Małgorzata, 10
Koziarz, Maria, 22
Kołodziejczyk, Karina, 26, 30
Krzywiecka, Anna, 27
Kupidura, Piotr, 28
Kurasińska, Amelia, 31
Kwiecińska, Julia, 22

Machnicka, Magdalena, 39
Merta, Julia, 13
Michalak, Agnieszka, 35
Mika, Justyna, 14
Muszer, Maja, 17
Młotkowski, Jakub, 17, 40

Najdek, Monika, 16, 18
Nowakowska, Alicja, 18
Nykiel, Katarzyna, 24

Ostašov,Pavel, 12

Patrycja Podleśny, 32
Pawłowska, Nikola, 13
Pietron, Damian, 36
Polańska, Oliwia, 10, 17, 25
Porzybót, Dominika, 20
Prudło, Agnieszka, 21
Pszenicyn, Michał, 26

Raczkowska, Patrycja, 34
Rakoczy, Paweł, 38
Rybarczyk, Nikola, 17, 25

Shyamsundar, Somanath, 36
Skutnik, Klaudia, 13
Stanowski, Michał, 33
Suwalska, Aleksandra, 23
Szczepińska, Teresa, 29
Szefczyk, Monika, 10
Szymik, Anna, 39
Szyszka, Martyna, 13

Tasinkiewicz, Anastazja, 21, 30
Tomkowski, Paweł, 18

Walski, Tomasz, 17
Wieromiejczyk, Olga, 27
Wilczyński, Bartosz, 39

Yakymenko, Daryna, 29
Ylla, Guillem, 27

Zdankiewicz, Maria, 40