BIT18
Book of Abstracts

28–30 June 2018, Toruń, Poland

PROGRAM COMMITEE:

- Prof. Wieslaw Nowak (*Nicolaus Copernicus University, Torun, Poland*)
- Prof. Jarek Meller (*University of Cincinnati, USA*)
- Prof. Jerzy Tiuryn (*University of Warsaw, Poland*)
- Dr. hab. Witold Rudnicki (*University of Warsaw, Poland*)

LOCAL ORGANIZING COMMITTEE:

- Prof. Wieslaw Nowak (*Nicolaus Copernicus University, Torun, Poland*)
- Dr. Aleksandra Gruca (*Silesian University of Technology, Gliwice, Poland*)
- Dr. Łukasz Pepłowski (*Nicolaus Copernicus University, Torun, Poland*)
- Dr. Jakub Rydzewski (*Nicolaus Copernicus University, Torun, Poland*)
- Dr. Katarzyna Walczewska-Szewc (*Nicolaus Copernicus University, Torun, Poland*)

Ministry
of Science
and Higher
Education
Republic of Poland

City of Toruń

Contents

Lectures

Deciphering the host/microbiome interaction: a Multi-Omics perspective

Cristian Coarfa[1,2]
[1] *Molecular and Cellular Biology Department,*
Baylor College of Medicine, Houston, Texas, USA
[2] *Dan L Duncan Comprehensive Cancer Center,*
Baylor College of Medicine, Houston, Texas, USA

Due to rapid methodological advances, the cost and expertise barrier for analysis of the microbiome has greatly advanced. Researchers can generate deep profiling of diverse body sites, furthering the understanding of complex diseases such as cancer cell and metabolic syndromes. However, profound advances often require creative integration with existing multi-omics profiles, including transcriptomics, proteomics, epigenomics. Approaches for scalable and effective integration of large of lipids data sets pose a significant challenge.

We present our framework for integration of microbiome data with complex host transcriptomics and phenomics profiles. In particular, we demonstrate an application to colorectal cancer, where data was collected via feces 16S rRNA metagenomics, gene panel transcriptomics from colon biopsies, and extensive dietary information. We demonstrate how finding gene/microbiome associations can uncover a focused and dysregulated transcriptomi profile associated with microbiome function, leading to complex applications such as drug discovery. In addition, associations between microbiome and diet in different cancer stages could inform further improvements in standards of care.

Next-Generation Onco-Panels – Opportunities and Challenges in Targeted Capture DNA and RNA Sequencing

David P Kreil[1]

[1] *Boku University Vienna*

With next-generation sequencing poised to enter routine clinical assays, the US FDA and other regulatory bodies have an interest in characterizing the measurement performance of this technology in a biomedical context. Targeted Capture of potential biomarkers of interest for subsequent sequencing allows the construction of comprehensive cancer panels to aid oncological diagnosis and prognosis. We here present recent emerging developments of complementing more traditional DNA based assays with probes for RNA, and discuss current and future opportunities and challenges in assessing not just single base variation but also larger scale rearrangements including gene fusions, as well as the longer term goal of quantitative profiling of alternative transcripts as a marker of direct oncological relevance

Intra- and inter-chromosomal chromatin interactions mediate genetic effects on regulatory networks

Alexandre Reymond[1]

[1] *Center for Integrative Genomics, University of Lausanne, Switzerland*

Population measurements of gene expression and genetic variation enable the discovery of thousands of expression Quantitative Trait Loci (eQTL), an extensive resource to determine the function of non- coding variants. To describe the effects of eQTL on regulatory elements such as enhancers and promoters, we quantified gene expression (mRNA) and three key histone modifications (H3K4me1, H3K4me3 and H3K27ac) across two cell types, 320 Lymphoblastoid Cell Lines and 80 Fibroblasts densely genotyped European samples. We find that nearby regulatory elements form local chromatin modules (LCM) often comprising multiple sub-compartments and overlapping topologically associating domains. These modules bring multiple distal regulatory elements in close proximity, vary substantially across cell types and drive co-expression of multiple genes. This regulation is under strong genetic control as 34,000 chromatin QTLs (cQTLs) affect 30% of the histone marks and as 70% of LCMs are associated with QTLs. These LCMs empower association studies of rare variants when whole genome sequencing is available. Using the Geuvadis transcriptomic data we unravel that expression of 20% of genes is associated with rare non-coding variants in modules for example. Coordination between regulatory elements located on different chromosomes (i.e. in trans) is well supported by Hi-C sequencing data and seem to drive in some cases trans eQTL effects. We replicated up to 80% of the strongest inter-chromosomal Hi-C contacts. Overall, this large-scale study integrating gene expression, chromatin activity and genetic variation across two cell types and hundreds of samples provides key insights into the biology underlying gene regulation and eQTLs.

Automated Inference of Gene Regulatory Networks Using Explicit Regulatory Modules

Clémence Réda[1] and Bartek Wilczyński[2]

[1]*École Normale Supérieure Paris-Saclay, Paris, France*
[2]*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland*

Gene regulatory networks are a popular tool for modelling important biological phenomena such as cell differentiation or oncogenesis. Efficient identification of the causal connections between genes, their products and transcription factors regulating them is key to understanding how defects in their function may translate to developmental defects or other diseases. Modelling approaches should keep up with the ever more detailed descriptions of the biological phenomena at play, as provided by new experimental findings and technical improvements. In particular, in recent years we have seen great improvements in mapping of specific binding sites of many transcription factors to distinct regulatory regions. Recent gene regulatory network models are to take advantage of the binding measurements, but usually only to define gene-to-gene interactions, ignoring regulatory module structure.

In our paper, we propose a method to specify possible regulatory interaction types in a given Boolean Network based on transcription factor binding evidence. This is implemented by an expansion algorithm that turns a regular Boolean Network model into a "Cis Regulatory" Boolean Network model, that explicitly defines regulatory regions as additional nodes in the network, and adds a new, valuable information to the dynamics of the network. The expanded model can be compared with expression data, and, for each node, a regulatory function consistent with the data can be found. The resulting models can then be inspected in in silico simulations.

The automated method for model identification has been implemented in Python, and the expansion algorithm in R. The method resorts to the Z3 SMT solver, and is similar to the RE:IN application.

It is available on https://github.com/regulomics/expansion-network.

Predicting cancer evolution from immune interactions

Marta Łuksza[1]
[1] *Icahn School of Medicine, Mount Sinai, New York*

In recent years, novel therapies for treating cancer by means of a patient's own immune system have emerged. Checkpoint-blockade immunotherapies are designed to enable immune system cells to recognize and destroy cancer cells. The process of recognition is based on specific protein binding interactions between the immune and cancer cells. Because these interactions depend on mutations in the cancer genome, immune recognition is also an evolutionary problem. I will present a mathematical model of cancer evolution based on the fitness cost of tumor cells due to immune recognition. The model successfully predicts response to checkpoint-blockade immunotherapy, as shown in patient cohorts with melanoma and lung cancer. Finally, the mechanistic model approach suggests principles for designing personalized treatments.

Whole exome and whole genome sequencing for discovery of novel human diseases

Rafal Ploski[1]

[1]*Department of Medical Genetics, Warsaw Medical University*

In 2012 Department of Medical Genetics (Warsaw Medical University) has acquired Illumina HiSeq 1500 which allowed to establish whole exome sequencing (WES) as method for both research and diagnostic purposes. Since then we have performed more than 1000 WES analyses, most of which aimed at finding diagnosis in patients suspected to suffer from rare disorders with a genetic basis. We also established shallow whole genome sequencing (WGS) based on Mate-pair libraries which we use to precisely map breakpoints in patients with symptomatic balanced chromosomal translocations. During the lecture selected findings will be presented illustrating how these approaches enable discovery of novel diseases (i.e. those caused by mutations in genes not yet associated with known human disorder).

Drug response prediction based on -omics data

Duc-Hau Le[1]

[1] *Thuyloi University, Hanoi, Vietnam*

Network biology: Large-scale data integration and text mining

Lars Juhl Jensen[1]
[1] *Novo Nordisk Foundation Center for Protein Research at the Panum Institute in Copenhagen, Denmark*

Methodological advances have in recent years given us unprecedented information on the molecular details of living cells. However, it remains a challenge to collect all the available data on individual genes and to integrate the highly heterogeneous evidence available with what is described in the scientific literature. The STRING database aims to address this by consolidating known and predicted protein–protein association data for a large number of organisms.

In my presentation, I will give an overview of the STRING database and describe the general approach we use to unify heterogeneous data, provide comparable quality scores for all evidence types, and automatically mine associations from the biomedical literature. I will also show how STRING and associated resources can be used through Cytoscape to visualize your own omics data on a network or to construct and compare, for example, disease-related protein networks.

Library of Integrated Network-based Cellular Signatures as a Resource for Precision Medicine

Jarek Meller[1]

[1] *Cincinnati Children's Hospital Medical Center and University of Cincinnati, Cincinnati, US*

LINCS library of cellular perturbation signatures is a unique resource that can be used to connect drugs and their targets, provide further insights into their mode of action, and open new avenues for personalized precision medicine. In this talk, an overview of LINCS datasets of transcriptional and proteomic signatures of chemical and genetic perturbations, as well as tools that are being developed to enable interactions with LINCS data, will be coupled with examples of use cases and applications. Collaborative projects involving basic and data scientists will be used to illustrate different strategies to generate specific hypotheses while connecting 'private' and 'public' data sets in the context of cancer, autoimmunity and brain disorders.

Pushing state-of-the art in transcriptomics and metagenomics on the road to personalized medicine

Pawel P Labaj[1, 2, 3]

[1] *Małopolska Centre of Biotechnology, UJ Kraków, Poland*
[2] *Austrian Academy of Sciences, Vienna, Austria*
[3] *Chair of Bioinformatics RG, Boku University Vienna, Austria*

In a more and more data driven society, services are increasingly tailored to better satisfy user specific needs [Dishman 2012]. In Precision or Personalized Medicine this means that prevention and treatment strategies take individual variability into account [Mirnezami 2012]. To identify this variability, we need to characterize individual healthy baseline of a person, in relation to population-based distributions [Shameer 2017]. This requires an improvement of measurement resolution as well as complementary new sources of biomedical and well-being data collected in real time [Milenković 2006, Bonaccorsi 2015]. These complementary data streams allow to build a comprehensive model of the response of person to internal (clinical) and external changes (exposome) and by this provide a detailed characterization of the healthy baseline. Such analyses subsequently allow the identification of true personal pathological changes. Detection of short-term episodes and long-term deviations from health enables early sensitive diagnosis while avoiding costly false calls [Ashley 2010, Chen 2012, Readhead 2013, Li-Pook-Than 2013]. In the talk I will present approaches how in my research I do address the challenges in both domains by pushing the state of the art in high-resolution clinical transcriptomics and characterizing the human exposome, applying metagenomics to the microbiome of the built-up environment in the context of personalized medicine

Ashley EA, et al. (2010) Lancet; 375: 1525-35.

Bonaccorsi, M., et al. (2015) Ambient Assisted Living: Italian Forum 2014. Springer. 465-475

Chen R, et al. (2012) Cell; 148 :1293-307.

Dishman (2012) [http://www.ey.com/gl/en/industries/life-sciences/the-personal-health-technology-revolution]

Li-Pook-Than, J. and Snyder M. (2013) Chem Biol; 20 :660-6.

Milenković, A., et al. (2006). Computer communications, 29(13), 2521-2533.

Mirnezami, R., et al. (2012) N Engl J Med; 366 :489-91.

Readhead, B. anf Dudley, J. (2013) Adv Wound Care; 2 :470-89.

Shameer, et al. (2017) Briefings in Bioinformatics; 18(1), 2017, 105-124

Accurate detection of clinically relevant variants using NGS data

Tomasz Gambin[1]

[1] *Warsaw University of Technology, Warsaw, Poland*

In this presentation, I will talk about two novel approaches to the analysis of NGS data, i.e. (i) accurate detection of small Copy Number Variants (CNVs), and (ii) the force calling approach to identify known disease Single Nucleotide Variants (SNVs) and short insertion/deletions (indels). In particular, I will describe main limitations of existing CNV callers and I will present a novel algorithm designed to identify single-exon, rare and intragenic homozygous and hemizygous deletions that may represent complete loss-of- function of the indicated gene. Next, I will explain why the force calling of known disease variants may improve the molecular diagnostic rate and help to detect mutations missed by standard NGS pipelines. Finally, I will present several examples of clinical cases that were solved thanks to the application of above mentioned tools.

Non-exponential peptide folding kinetics: modeling and experiment

Krzysztof Kuczera[1]

[1] *Departments of Chemistry and Molecular Biosciences,*
University of Kansas, Lawrence, KS, 66045, USA

Peptides and proteins tend to spontaneously form active native structures determined by their sequence. This folding process is a crucial biological phenomenon, and the disruption of correct folding often leads to disease. Interestingly, many small proteins exhibit single- exponential, or two-state folding, with a limited number of experimental examples of more complex cases. To help in understanding of the microscopic features of peptide and protein folding pathways, we focus on computer simulations and theoretical analysis of folding of WH21, a model helix-forming peptide. Based on a long-term molecular dynamic simulation, we perform a structural RMSD-based clustering. This allows for trajectory discretization, the assignment of sampled structures to ca. 200 clusters or microstates, determining their populations and a kinetic rate matrix describing transitions. To characterize the folding pathway we apply concepts of kinetic coarse-graining, optimal dimensionality reduction and transition path theory. This analysis shows that WH21 folding is highly non-exponential. Consistent models of low dimensionality, N=2, 3 and 4 are generated and their kinetic and structural features discussed. Overall, our work shows the complex, non-exponential nature of folding paths even in relatively simple systems.

Bioinformatics solutions for precision oncology: the MOBIT project

Miroslaw Kwasniewski[1]

[1] *Medical University of Bialystok, Bialystok, Poland*

Precision medicine largely relies on the possibility to use molecular markers that signal disease risk or presence before clinical signs and symptoms appear, and thus it is focused on prevention or early intervention rather than on the reaction at advanced stages of diseases. Consequently, many areas of medicine, including oncology, are moving towards tailored and individualized treatment for patients based on their molecular profiles and clinical characteristics. Nevertheless, precision medicine is relatively young and still growing field with various limitations, including narrow pool of effective prognostic and diagnostic markers or simplified solutions for patient's data interpretation, making decisions inadequate or not as personalized as expected. Therefore, to tackle these limits in the tailored diagnostics and therapy in oncology, the innovative project MOBIT (MOlecular Biomarkers for Individualized Therapy) was initiated. The aims of the project are: (i) the creation of software platform for collection, management, integration and analysis of -omics and clinical data, (ii) the development of software platform-assisted personalized diagnostic procedures based on integrated genomics, transcriptomics, proteomics, metabolomics and PET/MRI imaging data, and, (iii) the establishment of reference model for personalized tumor diagnosis and intervention in Non-Small Cell Lung Cancer as a model. The recent results and advantages of the MOBIT project will be presented.

Forensics intelligence through DNA analysis – how genomics is strengthening forensics

Wojciech Branicki[1, 2]
[1] *Malopolska Centre of Biotechnology JU, Krakow*
[2] *Central Forensic Laboratory of the Police, Warsaw*

Forensic DNA intelligence can be helpful when biological traces from a crime-scene are available but the investigators have no hypothesis concerning identity of the perpetrator. In such scenario, the forensic DNA phenotyping methods can be used to predict ancestry, appearance and age of this unknown individual and lead the investigation in a specific direction. Predictive DNA analysis has developed in forensics as a consequence of a general progress in DNA sequencing technologies and better understanding of the role of human genome variation in phenotype determination. In the forensic project NEXT we use the potential of massive parallel sequencing to discover new DNA predictors for appearance and ancestry and to develop prediction systems for accurate phenotype description of an unknown individual. The results will allow us to test and update predictive models for several appearance traits including age-related externally visible characteristics.

The study is supported by the National Centre for Research and Development in Poland, NEXT no. DOB-BIO7/17/01/2015.

G-protein-coupled receptors - novel structures and activation routes

Slawomir Filipek[1]
[1] *Univeristy of Warsaw, Warsaw, Poland*

G-protein-coupled receptors (GPCRs) mediate most of human physiological responses to external stimuli including neurotransmitters, hormones, and diverse environmental signals. GPCR ligands are highly diverse and include peptides, nucleotides, lipids, amino acids, and glycoproteins. As GPCRs are involved in many diseases and neurological disorders, they are targets for many presently used medicines and are of utmost interest for the development of novel therapeutic compounds. There is a rich diversity of available ligands for even single GPCR subtype. These ligands are often characterized as inverse agonists that suppress basal activity, full agonists that maximally activate the receptor, partial agonists that produce submaximal activity even at saturating concentrations, and the neutral antagonists that occupy the orthosteric binding site but do not affect basal activity. To complicate matters further, the efficacy of a ligand may depend on the downstream signaling pathway used to quantify activity (signaling via classical route using G protein or via newly discovered route using arrestin).

GPCRs are not simple ON/OFF machines but can exist in thousands of final states thereby leading the cell response in required direction[1]. Although this functional versatility is important for normal physiological signaling, it makes identifying effective therapeutics very challenging. Although there is little sequence homology among these receptors, the core structure of GPCRs is conserved and consists of seven transmembrane helices that are connected by intracellular and extracellular loops. The recent advances in GPCR crystallography provided about 30 structures of unique GPCRs, which can be used to make homology models of unknown receptors[2]. Since GPCRs represent the largest family of surface receptors, with approximately 800 members in human genome, the role of precise structure modeling and molecular dynamics simulations will be increasing.

[1] B. Trzaskowski, D. Latek, S. Yuan, U. Ghoshdastider, A. Debinski, S. Filipek, Action of molecular switches in GPCRs - theoretical and experimental studies, Curr. Med. Chem. (2012) 19, 1090-1109. doi: 10.2174/092986712799320556

[2] P. Miszta, P. Pasznik, J. Jakowiecki, A. Sztyler, D. Latek, S. Filipek, GPCRM - a homology modeling web service with triple membrane-fitted quality assessment of GPCR models, Nucleic Acids Res. (2018) in press. doi: 10.1093/nar/gky429

GPCRM: A New Tool for GPCR Modeling

Przemyslaw Miszta,[1] Pawel Pasznik,[1] Jakub Jakowiecki,[1] Agnieszka Sztyler,[1] and Slawomir Filipek[1]

[1]*Faculty of Chemistry & Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland*

The G-protein-coupled receptors (GPCRs) belong to a superfamily of cell signaling proteins, which 7 helices are embedded in the membrane. GPCRs have a crucial role in many physiological processes and in multiple diseases. Currently available drugs, many of which have excellent therapeutic benefits, target only a few GPCR members. Therefore, the need for new, high quality structures of GPCRs is enormous, but currently, only about 20% of receptors from that vast family of GPCRs were resolved with diffraction methods. To design new drugs it is necessary to determine 3D structures of GPCRs. The homology modeling service GPCRM[1–3] meets those expectations by greatly reducing the execution time of submissions (from days to hours/minutes) with nearly the same average quality of obtained models. The GPCRM service developed by BIOmodeling group at University of Warsaw to predict 3D structures of GPCRs, is one of few services employing homology modeling but GPCRM is continually upgraded in a semi-automatic way and the number of template structures has increased from 20 in 2013 to over 90 including structures the same receptor with different ligands which can influence the structure not only in the on/off manner. The final structure of a seven transmembrane helices bundle is generated employing multiple templates and profile-profile comparison while extra- and intra-cellular loops are reconstructed using MODELLER[4] and refined by ROSETTA[5]. Additionally, due to three different scoring functions (Rosetta, Rosetta-MP, BCL::Score) it is possible to select accurate models for the required purposes: the structure of the binding site, the transmembrane domain, or the overall shape of the receptor. The GPCRM service is still developed and nowadays contains many features which can be used by a broad range of researchers and students, modelers and experimentalists. The service is friendly enough to be used by inexperienced person when using automated mode.

[1] P. Miszta, P. Pasznik, J. Jakowiecki, A. Sztyler, D. Latek, S. Filipek, Nucleic Acids Research, 2018, https://doi.org/10.1093/nar/gky429

[2] gpcrm.biomodellab.eu

[3] D. Latek, P. Pasznik, T. Carlomagno, S. Filipek, PLOS ONE, 2013, 8(2), e56742

[4] M.Y. Shen, A. Sali, Protein Science, 2006, (15), 2507–2524

[5] A. A. Canutescu, R. L. Dunbrack, Protein Science, 2003, (12), 963–972

What do slipknotted membrane proteins hide?

Rafał Jakubowski,[1] Szymon Niewieczerzał,[1] and Joanna I. Sułkowska[1, 2]

[1]*Center of New Technologies, University of Warsaw, ul. Banacha 2c, Warsaw*
[2]*Faculty of Chemistry, University of Warsaw, ul. Pasteura 1, Warsaw*

The last two decades brought significant progress in recognition of presence of topologies such as knots, slipknots[1], lassos[2] and recently discovered links[3] in proteins. Therefore, among a variety of aspects investigated in biological systems to the date, we observe a growing interest in meaning of such topologies in life processes. While some mechanical properties of soluble slipknotted proteins have been already described[4], not much is known about membrane slipknotted proteins. Filling this gap, we present results of our multiscale investigations focused on a slipknotted membrane protein – BetP, a member betaine/choline/carnitine transporters family. This protein, considered as a model system for the signal transduction[5], consists three internal slipknots, accounting for a +3 1 ,4 1 ,+3 1 slipknot topology. With applied newly implemented implicit membrane model with a well known structure based model[6], we show that the unfolding pathways is highly affected by unexpected metastable states, what may lead to untypical lack of sequentiality in unfolding events scenario.

This work was supported by Ministry of Science and Higher Education grant, Ideas Plus II (0003/ID3/2016/64). Computational facilities of Interdisciplinary Centre for Modern Technologies, Nicolaus Copernicus University in Torun, Poland, are acknowledged.

[1] KnotProt: a database of proteins with knots and slipknots, M Jamroz, W Niemyska, EJ Rawdon, A Stasiak,KC Millet,P Sułkowski, JI Sulkowska, NAR, 2014, 10.1093/nar/gku1059

[2] LassoProt: server analyzing biopolymers with lassos. P Dabrowski-Tumanski, W Niemyska, P Pasznik, JI Sulkowska, NAR, 2016, 44, W383

[3] Topological knots and links in proteins, P Dabrowski-Tumanski, JI Sulkowska, PNAS, 2016, 114, 3415

[4] Jamming proteins with slipknots and their free energy landscape, JI Sulkowska, P Sułkowski, JN Onuchic, Phys. Rev. Lett. 2009 103 268103.

[5] The osmoreactive betaine carrier BetP from Corynebacterium glutamicum is a sensor for cytoplasmic K+, R Ruebenhagen, S Morbach, R Kraemer, EMBO J., 2001, 20, 5412

[6] Selection of optimal variants of Go-like models of proteins through studies of stretching JI Sulkowska, M Cieplak Biophys. J. 2008 95, 3174

ReMuS: a web-based tool for searching pathogenic variants in tissue-specific regulatory regions

Damian Skrzypczak,[1,2] Wojciech Fendler,[2,3] and Paweł Sztromwasser[2]

[1] *Wrocław University of Environmental and Life Sciences, Wrocław, Poland*
[2] *Department of Biostatistics and Translational Medicine,*
Medical University of Lodz, Łódź, Poland
[3] *Department of Radiation Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA*

Identification of pathogenic variants yields molecular diagnosis in 25-50% of patients with rare genetic disorders. Variant analysis is typically focused on protein coding genes hence causative mutations located in the non-coding regions of the genome could easily be missed. Vast size of the non-coding regions, great number of variants, and difficulty in predicting their impact on the phenotype are among the main reasons. New methods for filtering non-coding variants based on their relevance to the phenotype are needed.

We develop ReMuS, an on-line tool that will facilitate identification of regulatory regions potentially associated with the phenotype of a monogenic disease. Starting from a small set of genes implicated in the disease pathogenesis ReMuS finds regulatory features linked with these genes in several large scale repositories of tissue-specific genome-scale regulatory data. Customizable search and step-by-step process allows for iterative building of a tissue-specific set of regions that likely play a role in regulating expression of the input genes in the tissues affected by the disease. The growing inventory of regulatory data available in ReMuS includes at the moment coordinates of tissue-specific enhancers, transcription start sites, and regions of accessible chromatin from ENCODE and FANTOM5 repositories. Future releases of ReMuS will also enable inclusion of relevant binding sites of transcription factors and miRNAs, as well as locations of TAD boundaries adjacent to the input genes. ReMuS is in active development and will be made freely available in fall 2018.

Identification of single nucleotide variants associated with Alzheimer's disease accompanied by characterization of chromatin states in human brain

Marlena Osipowicz,[1] Marcelina Szczerba,[1,2] Hanna Kranas,[1]
Bartek Wilczyński,[1] and Magdalena A. Machnicka[1]
[1]*Institute of Informatics, Faculty of Mathematics,*
Informatics and Mechanics, University of Warsaw, Poland
[2]*Faculty of Biology, University of Warsaw, Poland*

Alzheimer's disease (AD) is a complex neurological disorder, for which the main risk factor is age. However, recent studies revealed several fully penetrant mutations, causative for the early-onset, familial form of AD and more than 20 genetic risk loci for the more frequent, late-onset, sporadic AD. Even though thousands of individuals have been analyzed by genome-wide association studies (GWAS), the arising genetic discoveries explain only a small proportion of AD heritability. Some part of this missing heritability may be revealed through whole-genome sequencing (WGS) but analysis of the resulting large datasets requires appropriate strategies for identification of relevant genomic variation.

Here we present a machine-learning approach for classification of healthy individuals and AD patients based on single nucleotide variants (SNVs) detected with WGS. Using the feature selection Boruta method (Kursa and Rudnicki 2010) and random forest classification method on WGS data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI) we were able to achieve accuracy rate of classification of 98,6% (+/- 1,3%). This result represents a huge advancement compared to previous attempts to solve the same classification problem with decision trees, based on GWAS and clinical data (56,08% classification accuracy (Erdoğan and Aydin Son 2014)). With our approach we identify around 6 000 disease-associated SNVs, which are enriched in coding regions of genes known to be associated with AD.

Moreover, we will discuss the possibility of using recent advancements in characterization of chromatin states in human brain to facilitate identification of functional, AD-associated variants in non-coding regions of the genome. Functional non-coding variants are much more difficult to identify than coding ones, mainly due to the fact that there are many more non-coding variants than the coding ones (usually less than 5% of identified variants in an individual are coding), but they can be important factors contributing to the development of complex neurological disorders, such as AD. We will describe possible applications of recently published chromatin contacts data from fetal brain (Won et al 2016) and information about location of active chromatin marks (histone modifications, open chromatin regions) in prioritization and annotation of non-coding variants.

This research was supported by Foundation for Polish Science grant POWROTY/2017-3/11. MAM was additionally supported by the Foundation for Polish Science (START fellowship). HK and BW were supported by Polish National Science Center SYMFONIA grant 2015/16/W/NZ2/00314.
Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

Posters

Use of machine learning and targeted next-generation sequencing in search of genetic risk factors of Alzheimer's disease

Marlena Osipowicz,[1] Marcelina Szczerba,[2] Bartek Wilczyński,[1] and Magdalena A. Machnicka[1]

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

Alzheimer's disease (AD) is a complex, heritable neurological disorder. Several fully penetrant mutations in the APP (amyloid precursor protein), PS1 (presenilin 1) and PS2 (presenilin 2) genes have been shown to be causative for rare early onset familial AD (EOAD). The more frequent late onset sporadic AD (LOAD) is considered to be polygenic with more than 20 genetic risk loci identified so far, mainly by genome-wide association studies (GWAS). However, these genetic factors explain only a minor fraction of AD cases. Our aim is to identify new genetic risk loci based on publicly available results of whole-genome sequencing (WGS) data and based on targeted next-generation sequencing of Polish AD patients.

To identify genetic loci which may contribute to the development of AD we have adapted a random forest classification method to discriminate between healthy individuals and AD patients based on Single Nucleotide Polymorphisms (SNPs) determined by WGS. We have used WGS results for 235 AD patients and 251 healthy control individuals provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI). In the first step we identified relevant disease predictors among 38 million SNPs detected in the WGS data using the Boruta method (Kursa and Rudnicki (2010) J. Stat. Softw.). Next, we built a random forest classifier based on 6 000 relevant SNPs. The classification accuracy, assessed using 10-fold cross validation, reached 98,6% (+/- 1,3%), suggesting that the identified set of SNPs may point to genomic loci highly relevant for AD heritability.

We expect that variation in non-coding regions of the genome may be an important risk factor for AD. To look for functional non-coding variants associated with EOAD we will conduct targeted next generation sequencing of regulatory regions active in human brain for a cohort of Polish EOAD patients. We will identify disease-associated variants based on a comparison of variant frequencies among patients and in a group of more than one hundred healthy elderly Polish control individuals. Finally, we will perform functional annotation of identified variants. To the best of our knowledge no systematic search for regulatory variants associated with EOAD has been carried out so far. Since known mutations can explain only around 5%-10% of the EOAD cases, results of this research can help to unravel the missing genetic etiology of this disease.

This research was supported by Foundation for Polish Science grant POWROTY/2017-3/11. MAM was additionally supported by the Foundation for Polish Science (START fellowship). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

Novel Posturographic Applications Using Virtual Reality Technologies

Beata Sokolowska,[1] Teresa Sadura-Sieklucka,[2] Ewa Sokolowska,[3] Dagmara Kabzinska,[1] Monika Gorecka,[1] Magdalena Lachwa-From,[1] Katarzyna Binieda,[1] Artur Kiepura,[1] Weronika Rzepnikowska,[1] Paweł Kowalczyk,[4] Anna Skrobisz,[5] Anna Sobanska,[6] Malgorzata Dylewska,[7] Michal Kolinski,[1] and Bogdan Lesyng[1, 8]

[1] *Mossakowski Medical Research Centre PAS, Warsaw, Poland*
[2] *Prof. E. Reicher National Institute of Geriatrics,
Rheumatology and Rehabilitation, Warsaw, Poland*
[3] *Faculty of Psychology, University of Warsaw, Poland*
[4] *The Kielanowski Institute of Animal Physiology and Nutrition PAS, Warsaw, Poland*
[5] *The Cardinal Stefan Wyszynski Institute of Cardiology, Warsaw, Poland*
[6] *Institute of Psychiatry and Neurology, Warsaw, Poland*
[7] *Institute of Biochemistry and Biophysics, Warsaw, Poland*
[8] *Faculty of Physics, University of Warsaw, Poland*

Virtual Reality (VR) according to a classical definition given by Sherman and Craig is "a medium composed of interactive computer stimulations that sense the participant's position and actions and replace or augment the feedback to one or more senses, giving the feeling of being mentally immersed or present in the stimulation (a virtual world)" [Sherman WR and Craig AB (2003): Understanding Virtual Reality. Interface, Application, and Design. Elsevier (USA) Publisher, p.38]. Novel and innovative VR technologies are widely applied in daily and professional life (computer games, sports, education, culture, advertisement or military) and in biomedicine. In particular, VR has been used in clinical settings as a training tool for surgeons and as a stimulation tool in cognitive or post-traumatic stress disorders, as well as in phobias or pain therapies. VR and force posturography are beneficial and efficient visual stimulation methods for the use in postural control studies. Postural instability and falls are common and devastating features of ageing and many neurological, vestibular, orthopedic/rheumatic disorders, or after ischemic stroke. The combination of VR technology and posturography has indicated new ways to comprehensive trainings and to test subject's postural stability.

The aim of this study is to evaluate posturographic tasks of training sessions applying a "Neuroforma" Virtual Reality system (http://www.neuroforma.pl/en/for-rehabilitation-centres/). The system is based on virtual complex games designed for several training algorithms.

Preliminary versions of posturographic training approaches and rehabilitation protocols/programs have been tested, compared and further developed. The VR posturography protocols might be helpful for individual and precision needs of patients with neuromuscular deficits/dysfunctions in the motor learning or rehabilitation. Novel solutions and applications will be presented.

Authors wish to acknowledge researches and students who participated in this project, in particular to Czerwosz L, Rakowski F from MMRCPAS, and to Błażejewska A, Charzewski Ł, Krzyśko K from UW. This study was carried out using the computational infrastructure of the Biocentrum-Ochota project (POIG.02.03.00-00-0030/09) and applying "Neuroforma" software of Titanis' technologies. Financial support was provided by MMRCPAS (Z-526/22).

Improving mapping with contigs assembly

Ania Macioszek[1] and Bartek Wilczyński[1]
[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw*

Experiments based on next generation sequencing, like ChIP-seq or RNA-seq, play a key role in modern biology. Usually they require bioinformatic analysis. In most cases, the essential step in the analysis is mapping sequenced fragments to the reference genome. The more the reference sequence differs from the actual sequence of analysed genome, the more errors may occur during this step, and hence, more potential errors can appear in downstream analysis. Due to variability between specimens and many other factors, a reference genome hardly ever represents the actual genome of a given individual; furthermore, obtaining the actual sequence is still a very challenging task. Here we propose an approach that allows to reduce the negative influence of the differences between reference and actual genome by defining a set of contigs and a remapping strategy based on contigs alignment.

Gene expression profiling of boar spermatozoa with differences in freezability

Chandra Pareek[1] and Leyland Fraser[2]
[1] *Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University*
[2] *Faculty of Animal Bioengineering, University of Warmia and Mazury in Olsztyn*

In this study transcriptome sequencing (RNA-Seq) was used to compare the expression gene profiles between spermatozoa RNA isolated from Polish large white (PLW) boars with good and poor semen freezability (GSF and PSF, respectively). Total RNA was isolated from spermatozoa of three GSF ($n = 3$) and three PSF ($n = 3$) boars, using a modified extraction procotocol1. RNA-Seq library preparation and paired-end sequencing on the Illumina NextSeq 500 platform were performed for transcriptome sequencing of the isolated RNA samples obtained from spermatozoa of the six boars. Using the DESeq2 pipeline, the gene expression profiling with RNA-Seq data identified a total of 18570 spermatozoa differentially expressed (DE) gene transcripts, including 9209 up-regulated and 9361 down-regulated transcripts. Among the identified transcripts, a total of 117 and 118 transcripts were identified as highly significantly ($p < 0.05$) up-regulated and down-regulated DE-gene transcripts, respectively. Furthermore, preliminary studies showed that several up-regulated DE-gene transcripts ($log2FC > 1$, $p < 0.05$) and down-regulated DE-gene transcripts ($log2FC < -1$, $p < 0.05$) were more noticeable in either freezability group. Comparison analysis of the gene expression profiles between boars of the freezability groups resulted in the identification of a total of 1220 significant DE-gene transcripts in both the GSF and PSF boars. Among the identified up-regulated gene transcripts ($n = 9209$), 2233 and 1696 DE-gene transcripts were identified in boars with GSF and PSF, respectively. Likewise, among the down-regulated gene transcripts ($n = 9361$), 1370 and 2844 DE-gene transcripts were detected in boars with GSF and PSF, respectively. Preliminary findings on the gene expression profiling of spermatozoa provided additional information on the sperm-related DE gene transcripts that might be relevance in study based on the freezability of boar semen.

Supported by a National Scientific Centre (NCN) project, Poland (2015/19/B/NZ9/01333).

Fraser L., Brym P., Pareek C.S. Isolation of total ribonucleic acid from fresh and frozen-thawed boar semen and its relevance in transcriptome studies. South African Journal of Animal Science, 2017, 47 (1): 56-60.

Chromatin compactness and gene expression level changes in D. melanogaster embryogenesis

Piotr Śliwa,[1] Aleksander Jankowski,[2] Yad Ghavi-Helm,[3] Eileen Furlong,[2] and Bartek Wilczyński[1]

[1] *Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland*

[2] *Genome Biology Unit, European Molecular Biology Laboratory (EMBL), D-69117 Heidelberg, Germany*

[3] *Institut de Genomique Fonctionnelle de Lyon, Univ Lyon,*
CNRS UMR 5242, Ecole Normale Superieure de Lyon,
Universite Claude Bernard Lyon 1, 46 allee d'Italie F-69364 Lyon, France

Throughout embryonic development we observe significant changes in gene expression levels - this allows the differentiation of the forming organism to occur. An important factor determining gene activity is the spatial organization of chromatin in the nucleus. Significant amount of regulation is mediated by enhancer-promoter interactions, level of compaction of the chromatin and its accessibility to transcription factors. We explore the spatial architecture of chromatin (HiC data) in relation to expression changes (RNAseq) in Drosophila melanogaster embryonic development in three timepoints (from undifferentiated to almost fully differentiated embryo).

Structural determinants of Kir6.2/SUR1 channel malfunctions related to diabetes – insights from genetics and bioinformatics

Katarzyna Walczewska-Szewc[1,2] and Wiesław Nowak[1,2]
[1] *Institute of Physics, Department of Biophysics and Medical Physics,*
Nicolaus Copernius University in Toruń, Poland
[2] *Interdisciplinary Centre for Modern Technologies,*
Nicolaus Copernius University in Toruń, Poland

Mutations in genes KCNJ11 and ABCC8 encoding the subunits of the ATP-sensitive potassium channel (K_ATP) may affect insulin release from pancreatic beta-cells. Both gain and loss of channel activity are observed, which leads to the varied clinical phenotype ranging from neonatal diabetes to congenital hyperinsulinism[1,2]. We would like to understand the mechanisms of the channel function impairments. To this end we mapped, based on the literature review, known medically relevant Kir6.2/SUR1 system mutations into recently (2017) discovered 3D structures of this complex. The assembly is composed of four inwardly rectifying K+ channel subunits (Kir6.2) and four sulfonylurea receptor (SUR1) moieties. Since the positions of several nucleotide binding sites have been determined, we may hypothesize about their role in the channel gating. Special role of PIP2 lipid in the insulin release process will be also discussed.

The authors acknowledge funding from National Science Centre, Poland (grant 2016/23/B/ST4/01770). The computational results used in this review were obtained using Interdisciplinary Centre for Modern Technologies, NCU.

[1] F.M. Ashcroft, M.C. Puljung, N. Vedovato, Trends Endocrinol Metab 28 (5), 377-387, 2017

[2] D. Ortiz, J. Bryan, Front Endocrinol (Lausanne) 6 (48), 2015

On the Origins of Life: Nano-confinement of Prebiotic Soup in Montmorillonite Clay – Car-Parrinello Quantum Dynamics Study

Juan Francisco Carrascoza Mayen,[1,2] Jakub Rydzewski,[3]
Natalia Szostak,[1,2,4] Jacek Blazewicz,[1,2,4] and Wieslaw Nowak[3]

[1]*Institute of Computer Science, Poznan University of Technology, Poland*
[2]*European Center for Bioinformatics and Genomics*
[3]*Institute of Physics, Faculty of Physics,*
Astronomy and Informatics, Torun, Poland
[4]*Institute of Bioorganic Chemistry, Poznan Academy of Sciences, Poland*

The catalytic effects of complex minerals or meteorites are often discussed as important for the origins of life. To assess the roles of a confinement and a strong surface electric field on the formation efficiency of simple precursors of nucleic acid bases or amino acids, we performed quantum Car-Parrinello molecular dynamics (CPMD) simulations. We prepared four condensed-phase systems modeled as prototypes of primordial soup. A montmorillonite clay (MMT) was used as a possible catalyst. We monitored chemical reactions in the MMT-confined simulation boxes on 20 ps time scale at 1 atm and 300 K, 400 K and 600 K. Elevated temperatures did not affect reactivity of elementary components, however, the presence of MMT substantially increased the formation probability of new molecules. The analysis of atom—atom radial distribution functions indicates that the presence of Ca+2 ions at the surface of internal cavities may be an important factor in initial steps of complex molecules formation at early stages of the Earth history and in those exo – planetary regions of the Universe where MMT type material is available.

dnaasm - de-novo assembler for second and third generation sequencing data

Robert Nowak,[1] Wiktor Kuśmirek,[1] Wiktor Franus,[1] and Mateusz Forc[1]

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

We propose a modification of the algorithm for DNA assembly, which uses the relative frequency of reads to properly reconstruct repetitive sequences (tandem repeats). The main advantage of our approach is that tandem repeats, which are longer than the insert size of paired-end tags, can also be properly reconstructed (other genome assemblers fail in such cases). What is more, tandem repeats could also be restored, if only single-read sequencing data is available.

We develop the application using bioweb skeleton, C++, Python, PostgreSQL and JavaScript, three-layered architecture, the data layer and the calculation layer is deployed on server site. The end user needs only web browser. The software was thoroughly tested, over 350 unit tests and about 25 simulated set of reads were used, almost 100% of code coverage was achieved. The results of the experiments presented proved the correctness of the algorithm and showed the effectively of the approach presented.

Additionally, we developed the scaffolding module, able to use long DNA reads with high level of errors, produced by third generation sequencers, like Oxford Nanopore. Therefore, our application allows to hybrid DNA assembly, where Illumina reads and Oxford Nanopore reads are used together to create longer contigs. The scaffolding module is based on Bloom filter and extremely memory-efficient associative array. Our implementation remarkably exceeds the previous one in terms of time and memory consumption.

Source code as well as a demo web application and a docker image are available at the dnaasm project web-page: `http://dnaasm.sourceforge.net`. The demo server is available at `http://eve.ii.pw.edu.pl:9007`

Hybrid de novo assembly of the Hymenolepis diminuta and Pichia fermentans genomes

Wiktor Kuśmirek[1] and Robert M. Nowak[1]

[1]*Institute of Computer Science, Warsaw University of Technology,*
Nowowiejska 15/19, 00-665 Warsaw, Poland

The aim of this study was to create reference genome for Hymenolepis diminuta (HD) and Pichia fermentans (PF).

For both organisms genetic material was isolated from DNA strains, then underwent Next Generation Sequencing using the Illumina HiSeq 1500 platformand a MinION sequencer from Oxford Nanopore Technologies (ONT) was done. For Illumina sequencer two types of data were obtained for HD genome: paired-end tags (mean insert size c.a. 400 bp) and mate-pairs (mean insert size c.a. 7 kbp). For PF genome we sequenced only paired-end tags (mean insert size c.a. 400 bp). For both organisms the MinION sequencing use standard protocol, where average read length was 10 kbp with 0.15 errors.

For HD genome obtained reads were de novo assembled in three steps. Firstly, paired-end tags were assembled by ABySS, Velvet and our dnaasm applications (http://dnaasm.sourceforge.net), the results from these applications were merged by GAM-NGS tool. The number of sequences longer than 1000 bp was equal to 6416, value of N50 - c.a. 70 kbp, sum - 166.120 Mbp. Secondly, DNA sequences were scaffolded using mate-pairs by SSPACE application. The N50 increased to c.a. 844 kbp, the number of sequences longer than 1000 bp decreased to 2342, and the sum of DNA sequences increased to 170.838 Mbp. Lastly, sequences obtained from paired-end tags and mate-pairs were joined using ONT sequencing data by LINKS tool. Final value of N50 was 2537 kbp, the number of sequences longer than 1000 bp decreased up to 719, and the sum of the sequences increased to 177.348 Mbp.

For PF genome reads were de novo assembled in two steps. Firstly, paired-end tags were assembled by dipSPAdes application. The number of sequences longer than 1000 bp was 511, value of N50 - c.a. 52 kbp, sum - 10.408 Mbp. Secondly, the assembler output was joined using ONT sequencing data by LINKS tool. Final value of N50 was equal to 134.5 kbp, the number of sequences longer than 1000 bp decreased up to 249 and the sum of the sequences increased to 10.649 Mbp.

In this study we showed, that merging data from different sequencing platforms could improve the final results of de novo assembling. Especially, combining short DNA reads obtained from next-generation sequencing with long DNA reads obtained from third generation sequencing could significantly improve the length and the quality of the resultant DNA sequences. Moreover, the increasing of coverage of the each sequencing technology separately, above the certain level, did not increase assembly results. Additionally, we proved that an assembly of highly heterozygous genome (PF) is much more complicated than homozygous genome (HD), despite the fact that the PF genome is more than 15 times smaller than HD genome.

HiCEnterprise: Identification of long-range interactions between chromatin regions

Hania Kranas[1] and Bartek Wilczyński[1]

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

Chromatin folding has an important role in bringing the distant regulatory elements in close proximity so that they can interact. HiCEnterprise is a toolkit for identification of such links, interactions between regions or domains, from chromosome conformation capture Hi-C data. The poster presents methods implemented in the package and provides an example use of HiCEnterprise for finding enhancer target genes and exploring higher-order chromatin structure.

The role of YRNA-derived small RNAs in atherosclerosis development and progression – modeled and analyzed using stochastic Petri nets

Agnieszka Rybarczyk[1, 2]

[1] *Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland*
[2] *Institute of Bioorganic Chemistry, Polish Academy of Sciences,*
Noskowskiego 12/14, 61-704 Poznań, Poland

Non-coding RNwAs are involved in a multitude of cellular processes and for many of them it has been demonstrated that they have a key role in regulating diverse aspects of development, homeostasis and diseases. Among them, YRNA-derived fragments are now of clinical interest and have attracted much recent attention as potential biomarkers for disease, since they are highly abundant in cells, tissues and body fluids of humans and mammals, as well as in a range of tumors.

In this study, to investigate the participation of the YRNA-derived small RNAs in the development and progression of the atherosclerosis, a stochastic Petri net model has been build and then analyzed. First, MCT-sets and t-clusters were generated, then knockout and simulation based analysis was conducted. The application of systems approach that has been used in this research has enabled for an in-depth analysis of the studied phenomenon and has allowed drawing valuable biological conclusions.

Integrative Galaxy tools for genomic data visualization - Genome Browsers and Hilbert Curve

Karolina Sienkiewicz[1] and Bartek Wilczyński[1]

[1] *Institute of Informatics, University of Warsaw*

Nowadays with the advance of high-throughput genomics and common application of machine learning to biological data, most of the current scientific projects are based on big data analysis. In such cases, deposition of raw experiment data in a public repository (for example GEO or ArrayExpress) is not sufficient and frequently supporting results with data visualization is desirable. Plenty of solutions were proposed to make a complex analysis of biological data more accessible, however, visualization of a large amount of data and sharing it remain an ongoing problem.

We are creating the set of integrated tools which allows exporting data from Galaxy web platform and visualizing it automatically in a variety of ways. Our main goal is to automatize data flow between computational biomedical research and data visualization for projects which conduct a genome analysis. This set of tools is going to incorporate diverse methods of genomic data visualization and data sharing: local visualization of genomic data on Hilbert Curve, automatic creation of Track Hubs for data visualization in UCSC Genome Browser and data visualization in individual instances of JBrowse Genome Browser.

Adaptive decomposition of ECG signals for anomaly detection using Orthogonal Matching Pursuit Algorithm

Sandra Śmigiel[1] and Damian Ledziński[1]

[1] *UTP University of Sciences and Technology*

In this paper, I present the use of adaptive decomposition method of the ECG signal for solving the problem of anomaly detection. The analyzed signal was presented as a set of correct ECG structures and anomalies (characterizing different types of disorders). In the course of adaptive decomposition I used the orthogonal matching pursuit algorithm and overcomplete Gabor dictionaries. In the process of anomaly detection based on decomposition of the analyzed signal onto projection coefficients, dictionary elements and residuals, was used energy approach. Performance of the proposed method was tested using a widely available database of ECG signals MIT– BIH Arrhythmia Database. The obtained experimental results confirmed the effectiveness of the method of anomaly detection in the analysed ECG signals.

Hierarchy of RNA folding reflected in DBL encoding

Maciej Antczak,[1,2] Mariusz Popenda,[2] Tomasz Zok,[3] Michal Zurkowski,[1,*] Ryszard Adamiak,[1,2] and Marta Szachniuk[1,2]

[1]*Institute of Computing Science, Poznan University of Technology*
[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences*
[3]*Poznan Supercomputing and Networking Center*

Contemporary structural biology puts a lot of effort to reveal and understand the relationship between structures and functions of biological molecules. RNA, being one of these molecules, attracts many scientists and it defines a research field with ever-widening borders. We already know that there is a variety of RNA types with different sizes, structures, and functions. However, it seems that there is much more waiting to be discovered.

Several observations have shown that the folding process of RNA molecule is hierarchical [Annu. Rev. Biophys. Biomol. Struct. 26, 1997,113-37]. Moreover, it seems that one RNA molecule can play various functions at different stages of folding. Thus, methods to trace, describe and show RNA folding hierarchy are needed. This necessity should be – at first – addressed by computational methods.

Here, we present two new algorithms to recognize pseudoknots in RNA structure. They are applied to (i) reveal an unknotted (core) RNA structure, (ii) find pseudoknots and (iii) classify them into subsets associated with pseudoknot orders. We follow the hypotheses stating that the core structure is formed first in the hierarchical folding, and pseudoknots occur in consecutive stages. We suggest that the order of pseudoknot corresponds to its formation order during the folding process. Our new algorithms are compared to the other methods that have existed before and touch the problem of pseudoknot processing. We show how the algorithms encode pseudoknotted RNA secondary structures in dot-bracket-letter (DBL) notation. In DBL, each pseudoknot order is represented by a different type of brackets or letters. Thus, DBL representation of RNA secondary structure shows not only where base pairs exist in the structure, but also at which stage of folding hierarchy they were created.

A detailed study of presented methods can be found in our recent paper [Bioinformatics 34(8), 2018, 1304-1312]. The algorithms have been implemented within RNApdbee 2.0 system [Nucleic Acids Res. 46(W1), 2018, in press, gky314] where they are freely available to everyone.

This work was supported by the National Science Center, Poland (2016/23/B/ST6/03931) and Faculty of Computing, Poznan University of Technology, within intramural financing program.

Analysis of the eQTLs and differentially expressed genes in the RNA-Seq data from glioma patients.

Ilona E. Grabowicz,[1, 2] Michał J. Dabrowski,[1] Bartosz Czapski,[3, 4] Tomasz Czernicki,[5]
Michał Draminski,[3] Bartłomiej Gielniewski,[3] Wiesława Grajkowska,[6] Bożena
Kaminska,[7] Katarzyna Kotulska,[7] Magdalena A. Machnicka,[8] Ania Macioszek,[9]
Paweł Nauman,[9] Karolina Stepniak,[3] Bartosz Wojtas,[3] and Bartek Wilczynski[8]

[1] *Computational Biology Lab, Institute of Computer Science, Polish Academy of Sciences*
[2] *Postgraduate School of Molecular Medicine, Medical University of Warsaw*
[3] *Laboratory of Molecular Neurobiology, Nencki Institute*
of Experimental Biology of Polish Academy of Sciences
[4] *Mazovian Brodno Hospital, Warsaw, Poland*
[5] *Medical University of Warsaw, Poland*
[6] *Children's Memorial Health Institute, Warsaw, Poland*
[7] *Children's Memorial Health Institute, Warsaw, Poland*
[8] *Computational Biology Group, Institute of Informatics, University of Warsaw*
[9] *Institute of Psychiatry and Neurology, Warsaw, Poland*

Gliomas represent 70% of brain tumours that according to World Health Organization classification are graded from I to IV. Glioblastoma, the most aggressive grade IV tumour, is characterized by median patients' survival from 15 months after diagnosis. In order to assess Single Nucleotide Polymorphisms (SNPs) as possible factors contributing to the pathobiology of gliomas, we have assessed their impact on the gene expression in 34 RNA-Seq samples coming from a cohort of Polish glioma patients representing all four grades. In total, we have identified 57 430 SNPs, out of which 163 correlated with gene expression (eQTLs). Two thirds of eQTLs were located within a gene body and one third nearby a gene. Majority of detected eQTLs (89%) correlated with up-regulated gene expression. Additionally, the expression of a subset of genes (n=13), with mapped eQTLs, correlated with the open chromatin marks: H3K4me3, H3K27ac and DNase I peaks determined by NGS sequencing. There were two other eQTLs that appeared to be localized within binding sites of transcription factors, namely: AREB6 and TEF-1. Moreover, we have identified the differentially expressed (DE) genes between different glioma grades (GI: Pilocytic Astrocytoma, GII/III: Diffuse Astrocytoma, GIV: Glioblastoma and Pediatric Glioblastoma) or between glioma and normal brain samples (DeSeq2, FDR p-val<0.01). Out of the identified eQTLs, 52 were mapped to DE genes. We found that DE genes were significantly enriched in genes known to be involved in cancerogenesis: oncogenes, tumour suppressors, epigenetic modifiers. Out of those three, the highest enrichment frequency was found for epigenetic modifiers among DE genes between GI and normal brain (hypergeometric distribution test, p-val = 5.4e-39). To summarise, we have identified genes which could be players in glioma formation. We have found genes whose expression changes might result from SNPs and they are potential candidates for further validation.

Pre-sequencing quality control of ATAC-seq libraries

Gosia Golda,[1,2] Lara Bossini-Castillo,[1] Natalia Kunowska,[1]
Dafni Glinos,[1] Pawel Labaj,[3,4,5] and Gosia Trynka[1]

[1] *Wellcome Sanger Institute, Cambridge, United Kingdom*
[2] *Faculty of Biochemistry, Biophysics, and Biotechnology, Jagiellonian University, Krakow, Poland*
[3] *Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland*
[4] *Austrian Academy of Sciences, Vienna, Austria*
[5] *Chair of Bioinformatics RG, Boku University, Vienna, Austria*

ATAC-seq was first described in 2013 and it was immediately embraced by the scientific community as a fast and straightforward assay for defining gene regulatory elements marked as chromatin accessible regions. Dependable results of the assay rely on the overall quality of ATAC-seq library. However, the clear-cut evidence for the quality of the experiment can be assessed only from the sequencing data. Having a reliable and comprehensive standard for estimating sample quality prior to sequencing could save money and help guide sequencing study design. Common practice for assessing the quality of the ATAC-seq library is to analyse the Bioanalyzer electropherograms. However, this approach can be inconsistent because of the subjective human interpretation. The goal of our study is to establish automated and reliable procedure for standardization of ATAC-seq quality control by devising the ATAC-seq Integrity Number (AIN) algorithm.

We propose a logistic regression model for more accurate estimation. Our dataset included 416 ATAC-seq libraries on: regulatory T cells, lymphoblastoid cell lines, macrophages and conventional T cells. Data was generated following original ATAC-seq and FAST-ATAC-seq protocol. The results showed better accuracy of the model (94.9%) over the human eye (56.5%) and its applicability across different cell types and protocols.

RNA Structure Annotation and Visualization with RNApdbee 2.0

Tomasz Zok,[1,2] Maciej Antczak,[1] Michal Zurkowski,[1] Mariusz Popenda,[3]
Jacek Blazewicz,[1,3] Ryszard Adamiak,[3] and Marta Szachniuk[1,3]

[1]*Institute of Computing Science, Poznan University of Technology*
[2]*Poznan Supercomputing and Networking Center*
[3]*Institute of Bioorganic Chemistry, Polish Academy of Sciences*

RNApdbee 1.0 was introduced in 2014 as a webserver to extract RNA secondary structure from 3D coordinates. It integrated several external tools into a coherent, unified pipeline. Additionally, RNApdbee was the first tool to address the issue of pseudoknot with the priority it needs. The original article defines a new quantitative feature – a pseudoknot order – and shows how it applies to RNA structures of various complexity and size.

Recently, we prepared an upgraded version of RNApdbee. It is now capable of parsing 3D coordinates in PDBx/mmCIF format – obligatory standard for all new submissions to PDB servers. Our adoption of this format includes preparation of adapters for external tools which are no longer officially maintained and therefore unable to read PDBx/mmCIF data.

We have also integrated additional tools – FR3D and R-chie – and largely updated our support for those present since the beginning. Furthermore, RNApdbee 2.0 is now capable of finding structural motifs in the input structure and extracting their coordinates in a format supported by RNAComposer – a fragment-based RNA 3D structure prediction method.

Finally, the pseudoknot order assignment problem has been addressed in a much more thorough way. We have developed several algorithms and benchmarked them to obtain a ranking. The top method is used by default in RNApdbee 2.0, however this is subject to user selection.

This work was partially supported by the National Science Center, Poland (2016/23/B/ST6/03931 and 2016/23/N/ST6/03779).

DNA methylation analysis of glioma samples: Comparison of DNA methylation patterns in gliomas of various grades or different IDH gene status

Agata Dziedzic,[1] Rafał Guzik,[2] Michał Draminski,[1] Bartosz Wojtas,[2] Karolina Stepniak,[2] Bartlomiej Gielniewski,[2] Bozena Kaminska,[2] Tomasz Czernicki,[3] Paweł Nauman,[4] Bartosz Czapski,[5] Wieslawa Grajkowska,[6] Katarzyna Kotulska,[6] and Michal J. Dabrowski[1]

[1] *Computational Biology Lab, Institute of Computer Science, Polish Academy of Science, 5 Jana Kazimierza St, 01-248, Warsaw, Poland*
[2] *Laboratory of Molecular Neurobiology, Nencki Institute of Experimental Biology of Polish Academy of Sciences, 3 Pasteur St, 02-093, Warsaw, Poland*
[3] *Medical University of Warsaw, 61 Zwirki i Wigury St, 00-001, Warsaw, Poland*
[4] *Institute of Psychiatry and Neurology, 9 Jana Sobieskiego St, 02-957, Warsaw, Poland*
[5] *Mazovian Brodno Hospital, 8 Ludwika Kondratowicza St, 03-242, Warsaw, Poland*
[6] *Children's Memorial Health Institute, 20 aleja Dzieci Polskich St, 04-730, Warsaw, Poland*

DNA methylation which is an epigenetic modification is important for understanding tumor initiation and progression. Methylation status of regulatory regions, such as promoters and enhancers, may activate or repress gene activity causing changes in gene expression levels. Recently, much evidence has been gathered which suggests that DNA methylation may play a crucial role in the glioma pathogenesis. Here we aimed to identify differentially methylated sites in respect to different tumor grades as well as IDH status. We investigated 21 brain tumour samples representing various glioma grades: pilocytic astrocytoma (PA; grade I; n = 7), diffuse astrocytoma (DA; grades II and III; n = 7) and glioblastoma (GB; grade IV; n = 7). We used SeqCap Epi CpGiant Methylation panel and performed bisulphite conversion followed by Illumina sequencing. For the same group of tumor samples RNA was isolated and the RNA-seq was performed. According to methylome data, we obtained at average 23 mln sites per each sample. From those 23 mln sites 3.3 mln were in CpG context and 19.7 mln were in non-CpG context. For further study we only selected sites that were in CpG context and with coverage above 10 reads. We performed differentional methylation calling and found 209,260 differentially methylated sites for PA vs. GB and 215,923 differentially methylated sites for IDH wild-type vs. IDH mutant. Then we annotated those sites in respect to the genome parts (promoter, gene body, CpG islands). Additionally, to verify the differentionally methylated sites functionality, we investigated expression (RNA-seq) of genes associated with those sites.

National Science Center, SYMFONIA3: DEC 2015/16/W/NZ2/00314

Analysis of transcriptomics data from mice faecal microbiome

Julia Herman-Izycka,[1] Ilona Grabowicz,[2] and Bartek Wilczynski[1]

[1] *Institute of Informatics, University of Warsaw*

[2] *Postgraduate School of Molecular Medicine, Medical University of Warsaw*

Microbiome, especially gut microbiome is known to influcence health of its host. Its composition is influenced both by diet and by host genetics.

Metatranscriptomics is a field investigating gene expression in population of microorganisms living in particular environment.

We analyze effect of diet change on a gut microbiome in two genetically different groups of mice via RNA-seq of faecal samples in order to find differentially expressed genes.

Application of the GWAS method in the analysis of the European bison (Bison bonasus) genome

Karol Puchała,[1] Zuzanna Nowak,[1] and Wanda Olech[1]

[1]*Department of Genetics and Animal Breeding,*
Warsaw University od Life Sciences, Warsaw, Poland

The study aimed to check the use of the GWAS method in European bison genome research. For the first time, the analysis for this species, was performed simultaneously for several traits. The analysis used a bovine microarray BovineHD BeadChip that contains 777,962 SNP markers. Seventeen European bison genotypes were subjected to the analysis of association. The number of markers, after the application of qualitative selection criteria, has been reduced to 17,405. The analysis was performed with use GenomeStudio 2.0 for qualitative selection, PLINK for association analysis and R environment for imaging of results on Manhattan plots. Level of MAF (minor allel frequency) was set at >0.005 (ten times less than in analysis performed on Bos Taurus). Level of call frequency was set at >0.85 (less than standard 0.9 use for GWAS). In the case of epididymal nodules, a small number of individuals and high stratification of the examined group made it impossible to obtain reliable results. Three disease entities were selected for the study of the genotype-trait coupling: pneumonia, nephritis and the presence of nodules on the epididymis. In the case of susceptibility to pneumonia, one highly significant marker was found near the SCN1A gene, for which a significant probability of coupling was demonstrated. A mutation within this gene can cause chronic pneumonia. Also, in the case of inflammation of the kidneys, three markers located near the PIK3C3 gene, whose overexpression, described in human, may lead to autoimmune diseases within the kidneys, have been identified. Markers obtained during analyses were included in the composition of the microarray dedicated special for European bison.

Motif search in promotor regions from patients diagnosed with glioma

Marta Jardanowska,[1] Magdalena A. Mozolewska,[2] Karolina Stepniak,[2, 3]
Magdalena A. Machnicka,[4] Michał J. Dabrowski,[5] and Bartosz Wilczyński[4]

[1]*Institute of Computer Science, Polish Academy of Sciences Warsaw, Poland, Both authors contributed equally to this work*

[2]*Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland, Both authors contributed equally to this work, Corresponding author m.mozolewska@ipipan.waw.pl.*

[3]*Nencki Institute of Experimental Biology, Warsaw, Poland.*

[4]*Institute of Informatics, University of Warsaw.*

[5]*Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.*

In recent years many researches have been focused on understanding the regulatory nature of genes resulting in glioma development. The main hypothesis is that specific transcription factors of the selected gene-promoter regions can distinguish the grade of the glioma tumor. This way we could find the transcription factors characteristic for the cancer grade, which will allow to adjust treatment to the patient. In this study, 326 DNA sequences, obtained by ChIP-seq experiment from Polish patients diagnosed with glioma representing all WHO grades, were used to predict the transcription-factors binding sites. We studied sequences of length +/- 2000 kb from Transcription Start Site (TSS) placed on active promoters with H3K4me3 mark, using two approaches: (i) known motif search PWMEnrich Bioconductor R package, and (ii) de novo method implemented in MEME suite. All analysis was made based on two databases: JASPAR and HOCOMOCO to widen the possible outcomes, and using two different backgrounds based on chromatin state: (i) active promoters in all samples, and (ii) inactive promoters in all samples. We searched for the motifs to identify transcription factors important in discern glioma grades. Hierarchical clustering of motifs dependent on the background showed the presence of different groups of motifs resulting from different nucleotide enrichment levels.

Supported by a by a grant from the Polish National Science Centre [DEC-2015/16/W/NZ2/00314].

S.R. and D. D, PWMEnrich: PWM enrichment analysis. R package version 4.6.0., (2015).

T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers., Proceedings. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28–36. http://www.ncbi.nlm.nih.gov/pubmed/7584402 (accessed May 21, 2018).

A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J.A. Castro-Mondragon, R. van der Lee, et al., JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework, Nucleic Acids Res. 46 (2018) D260–D266. doi:10.1093/nar/gkx1126.

I. V Kulakovskiy, I.E. Vorontsov, I.S. Yevshin, R.N. Sharipov, A.D. Fedorova, E.I. Rumynskiy, et al., HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis, Nucleic Acids Res. 46 (2018) D252–D259. doi:10.1093/nar/gkx1106.

Euler angles in n-way junction modeling and analysis

Jakub Wiedemann,[1] Maciej Milostan,[1,2] Marta Szachniuk,[1,2] and Ryszard W. Adamiak[1,2]

[1] *Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology*
[2] *Institute of Bioorganic Chemistry, Polish Academy of Sciences, Z. Noskowskiego 12/14, 61 704 Poznan, Poland*

In the last decade, high scientific activity could be observed in the field of a development of computational algorithms for modeling and analysis of RNA three-dimensional structures. However, the prediction methods of 3D structures are still far from perfect. Current prediction methods are successful in handling quite many structure elements, but some motifs are not yet modelled in a reliable way. N-way junction (with $N > 2$) is an example of structure motif that is found hard to predict accurately by most computational algorithms.

In our work, we have collected all n-way junction structures found in experimentally determined RNAs and we analyzed their features. The motifs were identified using own search algorithm operating on dot-bracket representations of the input structures. The junctions were gathered to create the new n-way junction repository. For each candidate, a digraph model was proposed to represent selected features of its secondary structure and values of Euler angles describing the direction of outcoming stems. We believe this data can be used in the process of modeling of unknown RNA 3D structures and in the refinement of the existing ones.

This work was supported by the National Science Centre, Poland [2016/23/B/ST6/03931].

Bioinformatic Approach to Visualization and Analysis of G-quadruplex Structures

Joanna Miskiewicz,[1] Mariusz Popenda,[2,1] Joanna Sarzynska,[2,1]
Maciej Antczak,[1] Tomasz Zok,[1] and Marta Szachniuk[2,1]

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

G-quadruplexes are non-canonical structural formations that exist in guanosine-rich DNA, RNA and even in nucleic acids analogs. The G-quadruplex form may be built by one, two or four strands and their orientation determine the polarity of the G-quadruplex structure – parallel, antiparallel, and hybrid-type antiparallel. Due to a specific structure of G-quadruplex, it is involved in the various biological processes, such as mRNA processing, regulation, and transcription, which may be influenced by recruiting protein factors. Moreover, G-quadruplex structures are a promising target in many strategies of drug development, including anticancer and neurological disease therapies.

In our research, we use bioinformatic approach to visualize and analyze G-quadruplex structures in human microRNAs (small non-coding molecules). We propose a new manner to visualize their secondary structure in all possible combinations of triple G-tetrads. From PDB database we gathered structures recognized as the ones containing G-quadruplexes, both from RNA and DNA. We visualized these structures and defined their strands orientation in G-quadruplexes by using RNApdbee program. Based on these information, we categorize each G-quadruplex structure to one of the 24 proposed representations.

This work was supported by the National Science Centre, Poland [2016/23/B/ST6/03931].

Rhodes D., Lipps HJ., G-quadruplexes and their regulatory roles in biology, Nucleic Acids Research, 2015, 8627-8637

Malkowska M., Czajczynska K., Gudanis D., Tworak A., Gdaniec Z., Overview of the RNA G-quadruplex structures, Acta Biochimica Polonica 2016, 609-621

Fay MM., Lyons SM., Ivanov P., RNA G-quadruplexes in Biology: Principles and Molecular Mechanisms, Journal of Molecular Biology 2017, 2127-2147

Berman HM., Westbrook J., Feng Z., Gilliland G., Bhat TN., Weissig H., Shindyalov IN., Bourne PE., The Protein Data Bank, Nucleic Acids Research 2000, 235-242

Antczak M., Zok T., Popenda M., Lukasiak P., Adamiak RW., Blazewicz J., Szachniuk M., RNApdbee – a webserver to derive secondary structures from pdb files of knotted and unknotted RNAs. Nucleic Acids Research 2014, W368-W372

Motif search in promotor regions from patients diagnosed with glioma

Marta Jardanowska,[1, 2] Magdalena A. Mozolewska,[1, 3, 2] Karolina Stepniak,[4]
Magdalena A. Machnicka,[5] Michał J. Dabrowski,[1] and Bartosz Wilczynski[5]
[1] *Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland*
[2] *Both authors contributed equally to this work*
[3] *To whom the correspondence should be addressed: m.mozolewska@ipipan.waw.pl*
[4] *Nencki Institute of Experimental Biology, Warsaw, Poland*
[5] *Institute of Informatics, University of Warsaw*

In recent years many researches have been focused on understanding the regulatory nature of genes resulting in glioma development. The main hypothesis is that specific transcription factors of the selected gene-promoter regions can distinguish the grade of the glioma tumor. This way we could find the transcription factors characteristic for the cancer grade, which will allow to adjust treatment to the patient. In this study, 326 DNA sequences, obtained by ChIP- seq experiment from Polish patients diagnosed with glioma representing all WHO grades, were used to predict the transcription-factors binding sites. We studied sequences of length +/- 2000 kb from Transcription Start Site (TSS) placed on active promoters with H3K4me3 mark, using two approaches: (i) known motif search PWMEnrich Bioconductor R package, and (ii) de novo method implemented in MEME suite. All analysis was made based on two databases: JASPAR and HOCOMOCO to widen the possible outcomes, and using two different backgrounds based on chromatin state: (i) active promoters in all samples, and (ii) inactive promoters in all samples. We searched for the motifs to identify transcription factors important in discern glioma grades. Hierarchical clustering of motifs dependent on the background showed the presence of different groups of motifs resulting from different nucleotide enrichment levels.

Supported by a by a grant from the Polish National Science Centre [DEC- 2015/16/W/NZ2/00314].

S.R. and D. D, PWMEnrich: PWM enrichment analysis. R package version 4.6.0., (2015)

T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers., Proceedings. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28–36. http://www.ncbi.nlm.nih.gov/pubmed/7584402 (accessed May 21, 2018)

A. Khan, O. Fornes, A. Stigliani, M. Gheorghe, J.A. Castro-Mondragon, R. van der Lee, et al., JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework, Nucleic Acids Res. 46 (2018) D260–D266. doi:10.1093/nar/gkx1126

I. V Kulakovskiy, I.E. Vorontsov, I.S. Yevshin, R.N. Sharipov, A.D. Fedorova, E.I. Rumynskiy, et al., HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis, Nucleic Acids Res. 46 (2018) D252–D259. doi:10.1093/nar/gkx1106.

GBSC: Graph based method for clustering of low complexity regions

Patryk Jarnot,[1] Joanna Ziemska,[2] Marcin Grynberg,[3] and Aleksandra Gruca[1]
[1] *Institute of Informatics, Silesian University of Technology, Poland*
[2] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*
[3] *Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Poland*

Low complexity regions (LCRs) are abundant in the protein world. About 14% of proteins contain LCRs[1]. They sometimes play key roles in protein functions and are relevant to protein structures[2–4]. Therefore, the ability to identify similar LCRs in different protein sequences could provide the answer to the question if similar low complexity fragments share similar biological function. However, current methods for searching for similarities in protein sequences are not designed for analysis of low complexity regions, and some of them even mask LCRs to improve searching for homologous high complexity proteins. In order to overcome this problem we propose a new method designed specifically for searching for similar low complexity regions.

We present the graph based sequence clustering (GBSC) method, a new approach to analysis of LCRs. The proposed algorithm builds a graph from a sequence and then retrieves cycles. Nodes are created from k-mers of the sequence. The positions of each related k-mer are tagged and transitions connect neighboring k-mers. Each cycle which occurs more often than a given threshold represents a cluster, so one sequence can be assigned to more than one cluster. Cycles in a graph reflect repetitions of the sequence. Our method also handles fused LCRs, that is sequences composed of more than one type of repeats. When clustering of repeats is finished GBSC starts the next iteration. For each pair of unclustered sequences, the method joins their corresponding graphs by merging the same nodes and subsequently calculates the similarity of sequences. As a result, LCRs which do not contain repeats are also clustered.

[1] Marcotte E., Pellegrini M., Yeates T., Eisenberg D. *A census of protein repeats.* J Mol Biol. 1999;293:151–160.

[2] Hamm D.C., Bondra E.R., Harrison M.M. *Transcriptional activation is a conserved feature of the early embryonic factor zelda that requires a cluster of four zinc fingers for DNA binding and a low-complexity activation domain.* The Journal of Biological Chemistry. 2015;290(6):3508-3518.

[3] O'Rourke T.W., Loya T.J., Head P.E., Horton J.R., Reines D.*Amyloid-like assembly of the low complexity domain of yeast Nab3.* Prion. 2015;9(1):34-47.

[4] Kumari B., Kumar R., Kumar *Low complexity and disordered regions of proteins have different structural and amino acid preferences.* Mol Biosyst. 2015 Feb;11(2):585-94.

BEDroom: Python framework for building web-based applications operating on categorized sets of BED files

Damian Skrzypczak[1, 2]

[1] *Wrocław University of Environmental and Life Sciences, Wrocław, Poland*
[2] *Department of Biostatistics and Translational Medicine, Medical University of Lodz, Łódź, Poland*

Browser Extensible Data format (BED) is flexible and versatile way for storing practically every positional, genome-related data. Keeping mapped and annotated genetics abstractions (e.g. genes, promoters, or introns) as mathematical intervals, enables alignment of many layers of data in order to determine their relative positions (e.g. overlapping of known genes with regions of high conservation, or regulatory elements active in a given tissue). This in turn allows asking biologically relevant questions and is foundation of many bioinformatic and statistical analyzes.

Several software applications for manipulation of BED-formatted data exist. However, these tools are often designed for programmers, encapsulated as programming language libraries/modules or with command line interface. Meeting needs of users without programming expertise requires easier access and applications based on web infrastructure are a common way of providing user-friendly interfaces to many bioinformatics tools..

We are developing BEDroom, a Python framework for building web-based applications operating on categorized sets of BED files. BEDroom will provide a default application template, requiring only definition of the interface structure, data source and links between user actions and corresponding server-side operations. At the same time, the platform will be fully modifiable and expandable for the diverse needs of application developers. Requiring minimal effort in designing the user interface, BEDroom provides tools for reading collections of BED files organized in various ways, e.g. flat or tree structure of directories with BED files, databases, or single file with standalone records.

BEDtools is developed under MIT license and will be available soon at https://github.com/DamianSkrzypczak/ and as a Python library.

MotifLCR: Pattern and PSSM based method for clustering low complexity regions

Joanna Ziemska,[1] Patryk Jarnot,[2] Aleksandra Gruca,[2] and Marcin Grynberg[3]

[1] *Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*
[2] *Institute of Informatics, Silesian University of Technology, Poland*
[3] *Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Poland*

Low complexity regions (LCRs) can be found in proteomes more often than by chance. About 14 of proteins contain LCRs[1] and these may often be crucial for structure and function of proteins[2–4]. One can find a plethora of methods to find LCRs, however there is a serious scarcity of programmes to compare LCRs. In order to overcome this problem, we propose a new method designed specifically for searching similar repeat low complexity regions in diverse protein sequences.

This new algorithm finds repeats in sequences. It uses sliding window in order to determine the longest repeat. Each analyzed repeat cannot contain subrepeats. After assigning sequences to clusters basing on the same repeats the algorithm builds PSSMs that represent specific clusters. If clusters have similar PSSMs then these are merged. This method detects only consecutive repeats, with no insertions allowed in between. Substitutions and insertions are the most frequent mutations in repeats, whereas deletions are extremely rare. Most of these 'noises' in input sequences are removed before processing. Sequences not assigned to any cluster will be compared with existing clusters and assigned to most similar ones. If the given threshold is not met then this specific LCR will form a new orphan cluster. MotifLCR also creates clusters of LCRs composed of two or more fused low complexity fragments.

[1] Marcotte E., Pellegrini M., Yeates T., Eisenberg D. *A census of protein repeats.* J Mol Biol. 1999;293:151–160.

[2] Hamm D.C., Bondra E.R., Harrison M.M. *Transcriptional activation is a conserved feature of the early embryonic factor zelda that requires a cluster of four zinc fingers for DNA binding and a low-complexity activation domain.* The Journal of Biological Chemistry. 2015;290(6):3508-3518.

[3] O'Rourke T.W., Loya T.J., Head P.E., Horton J.R., Reines D.*Amyloid-like assembly of the low complexity domain of yeast Nab3.* Prion. 2015;9(1):34-47.

[4] Kumari B., Kumar R., Kumar *Low complexity and disordered regions of proteins have different structural and amino acid preferences.* Mol Biosyst. 2015 Feb;11(2):585-94.

Atherosclerosis process analyzed with stochastics Petri net model and its simulation

Marcin Radom,[1, 2] Dorota Formanowicz,[3] and Piotr Formanowicz[1, 2]

[1]*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland*
[2]*Institute of Bioorganic Chemistry, Polish Academy of Sciences,*
Noskowskiego 12/14, 61-704 Poznan, Poland
[3]*Department of Clinical Biochemistry and Laboratory Medicine,*
Poznan University of Medical Sciences, Rokietnicka 8, 60-806 Poznan, Poland

The complexity of atherosclerosis process makes it particularly difficult to fully understand; thereby having its detailed model greatly contribute to the discovery of new facts about this complex phenomenon. The whole process involves many different molecules and chemical compounds, as well as large number of various reactions that directly and indirectly interact with each other. Formation of an atherosclerotic plaque is a long lasting process with different dynamics, depending on the microenvironment in a blood vessel (in which can coexist lipids disorders, chronic inflammation, oxidative stress, shear forces and immune reponse). Due to this complexity the analysis of a single phenomenon (i.e. subprocess) is not sufficient and it can lead to false conclusions. To take into account many of the coexisting dependencies, there is a need for a systems approach. For this purpose a stochastic Petri net based model of atherosclerosis has been created. Its studying involves the structural analysis (based on invariants, MCT sets and clusters) but more importantly the simulation of the behavior of the model. Such simulation allows to acquire new insights about the influence of the process basic components on each other and on the atheroscleoris progression in the whole biological system.

This research has been partially supported by the National Science Centre (Poland) grant No. 2012/07/B/ST6/01537.

A mathematical approach for risk of death estimation due to cardiovascular diseases in Poland based on Pol-SCORE 2015 tables

Jakub Olszak,[1,2] Dorota Formanowicz,[3] and Piotr Formanowicz[1,4]
[1]*Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland*
[2]*Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Holbeck, Leeds LS11 5PY, U.K*
[3]*Department of Clinical Biochemistry and Laboratory Medicine,*
Poznan University of Medical Sciences, Rokietnicka 8, 60-806 Poznan, Poland
[4]*Institute of Bioorganic Chemistry, Polish Academy of Sciences,*
Noskowskiego 12/14, 61-704 Poznan, Poland

A SCORE (Systematic COronary Risk Evaluation) is a set of high and low cardiovascular risk charts based on a large dataset based on the European population. They are dependent on gender, age, total cholesterol, systolic blood pressure and smoking status. The Polish population was initially classified by high-risk charts, but the latest studies showed that the SCORE 2007 function overestimated cardiovascular risk in Poland. According to these studies, Pol-SCORE 2015 charts were presented and reflected a more accurate status of the current generation.

In this work a mathematical model will be proposed to calculate the risk of cardiovascular disease. It is based on Pol-SCORE 2015 charts and reflects the dependencies between all of the parameters which affect the risk of cardiovascular disease. From this model it is possible to calculate the estimated risk for a wider range of parameters or alternatively without the knowledge of actual chart values. The models accurateness for the non-smoking population of men in the worst case varies around 2-3% compared to the values taken from the Pol-SCORE 2015 charts.

This research has been partially supported by the National Science Centre (Poland) grant No. 2012/07/B/ST6/01537.

Tomasz Zdrojewski, Piotr Jankowski, Piotr Bandosz et al. Nowa wersja systemu oceny ryzyka sercowo-naczyniowego i tablic SCORE dla populacji Polski. Kardiologia Polska 2015, 73, 958–961.

A systems biology approach to study of different pathways of interleukin 18 synthesis

Kaja Chmielewska,[1] Dorota Formanowicz,[2] and Piotr Formanowicz[1,3]
[1]*Institute of Computing Science, Poznan University of Technology*
[2]*Department of Clinical Biochemistry and Laboratory Medicine, Poznan University of Medical Sciences*
[3]*Institute of Bioorganic Chemistry, Polish Academy of Sciences*

Interleukin 18 (IL-18) is a pleiotropic pro-inflammatory cytokine recognized as very important regulator of innate and adaptive immune responses. Moreover, IL-18 is involved in the development of various cardiovascular diseases, however, there is no definitive evidence of an accurate IL-18 action mechanism. Recent study suggest, that IL-18 can be treated as a mediator in the pathogenesis as well as a diagnostic marker. In view of showing the importance of interleukin 18 a Petri net-based model has been developed. In the proposed model two different pathways of IL-18 synthesis has been distinguished. The first one is caspase-1-dependent pathway and the second one is caspase-1-independent pathway. The analysis of models expressed in the language of Petri nets theory can be based on t-invariants, which correspond to subprocesses occuring in the modeled system. Searching for similarities between t-invariants may lead to identification of sub-processes which can influence each other. The analysis focused on IL-18 allows to determine that this pro-inflammatory cytokine is produced more often via caspase-1-independent pathway than caspase 1-dependent pathway. Furthermore, it appears that caspase 8 may be associated with caspase-1-independent pathway. This discovery is consistent with new research results[1].

This research has been partially supported by the National Science Centre (Poland) grant No. 2012/07/B/ST6/01537.

[1] G. Kaplanski: Interleukin-18: Biological properties and role in disease pathogenesis. Immunol Rev. (2018), 281(1), 138-153.

Simulation of chromatin structure

Irina Tuszynska,[1] Pawel Bednarz,[1] and Bartek Wilczynski[1]

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

The genomic DNA is a part of the chromatin, which forms in the cell nucleus a complex spatial structures necessary for the proper functioning of the cell in the interphase. Chromatin should form of particular spatial structure to effect certain processes such as replication and repair of DNA or transcriptional regulation.

Experimental methods, such as FISH, Chia-PET, Hi-C and 3C, 4C, 5C, revealed a great multitude of structures that chromatin adopted in different regions of the genome, different cell types and in different environmental conditions . These methods, however, do not allow to directly reveal the structure of chromatin - they show only the average frequency of contacts of individual regions of chromatin. Theoretical methods are extremely useful and effective tool to support experimental research. Here we have used the SBS model with experimental data of APBS (architectural protein binding sites) to predict structure of Drosophila melanogaster chromosomes. We compared predicted structures of chromatin with experimental data (Hi-C maps). The similarity of fourth chromosome structure contact map with it's Hi-C map encourages us to apply the same method for prediction the structure of all chromosomes of Drosophila melanogaster.

Transcriptome sequencing of liver from Polish Landrace and Polish Landrace x Duroc pigs in response to omega-3 fatty acids action

Agnieszka Szostak,[1, 2] Magdalena Ogluszka,[2] Marinus F. W. te Pas,[3] Ewa Polawska,[2] Pawel Urbanski,[2] Tadeusz Blicharski,[2] Jaroslaw O. Horbanczuk,[2] and Mariusz Pierzchala[2]

[1] *Department of Genetics and Animal Breeding,*
Warsaw University of Life Sciences, Warsaw, Poland
[2] *Institute of Genetics and Animal Breeding,*
Polish Academy of Science, Jastrzebiec, Poland
[3] *Animal Breeding and Genetics Centre, Wageningen UR Livestock Research, Lelystad, The Netherlands*

Background: Omega-6 and omega-3 polyunsaturated fatty acids (PUFAs) have been recognized as molecules regulating variety of functions in the cell. They serve as a source of energy, are a vital component of cell membranes and act as signaling molecules, which can regulate gene expression. The debate over the impact of omega-6 and omega-3 PUFAs on potential health outcomes has been settled, however it is still not definitively clear what is the mode of action on the whole transcriptome level. The aim of the study was to investigate the effects of dietary omega-6 and omega-3 fatty acids on hepatic genome activity in pigs utilising transcriptomic analysis. Additionally under our consideration were the alterations in gene expression associated with genotype, examined on Polish Landrace pure breed and Polish Landrace x Duroc crossbreed pigs. Methods: The animals were fed a diet enriched with omega-6 and omega-3 PUFAs or a standard diet. The hepatic profiles of fatty acids were analyzed by gas chromatography (GC-FID) ($n = 160$). Transcriptome profiling was performed with Illumina apparatus using an RNA-sequencing approach (RNA-seq). Differentially expressed genes (DEGs) were identified using the CLC Genomics Workbench v. 6.0. The functional analysis was performed using the Database for Annotation, Visualization, and Integrated Discovery (DAVID v. 6.7) and the protein-protein interaction networks were constructed using Cytoscape v. 3.1.0 software. The results were confirmed and validated on larger groups of animals using real-time PCR ($n = 40$/genotype).

Results: Fatty acid profiles as determined by a gas chromatography protocol confirmed a fundamental role of omega-6 and omega-3 fatty acids in homeostasis, through markedly improved hepatic lipid composition in both of the case groups. Performing RNA-seq, we found 3584 DEGs within purebreed groups, and 4502 within crossbreed groups; 589 of which were common between the two genotypes. Differential expression of genes was considered significant at $P<0.05$, and a false discovery rate of<0.05 was implemented, with a fold change of $1.2 - 4$. Decreased omega-6/omega-3 ratio in the liver affected: a set of genes related to enhancement of fatty acid beta-oxidation, lipid catabolic processes, lipid transport, cholesterol efflux and inhibition of autophagic (specifically, lipophagic) processes in both genotypes belonging to the experimental class. Additionally, dietary omega-6 and omega-3 fatty acids influenced gene expression in divergent manners depending on the genotype, affecting either the immune system (Polish Landrace x Duroc) or signaling processes (Polish Landrace).

Conclusions: Dietary PUFAs are known to exert health promoting effects and can be used as a therapeutic strategy against deleterious lipid accumulation in the liver. The mechanism of such is through an overall improvement of lipid profile. Collectively, pathways elucidated in our study suggest that omega-3 fatty acids decrease cellular lipid accumulation (triglyceride content as lipid droplets), enhance cholesterol efflux outside the hepatocytes, and downregulate lipophagic processes.

The research was realized within the project BIOFOOD: innovative, functional products of animal origin; no. POIG.01.01.02-014-090/09 co-financed by the European Union from the European Regional Development Fund within the Innovative Economy Operational Programme 2007; 2013, and this work was supported by The National Science Centre, grant no. 2014/13/N/NZ9/00906.

Modelling the dynamics of interactions between transcription factors, stress response elements and gene expression patterns in cellular response to stress

Aleksandra Cabaj,[1] Agata Charzynska,[1] Rafal Bartoszewski,[2] and Michal Dabrowski[1]

[1] *Laboratory of Bioinformatics, Nencki Institute of Experimental Biology of Polish Academy of Sciences, Warsaw*

[2] *Department of Biology and Pharmaceutical Botany, Medical University of Gdansk, Gdansk*

One of the most important questions in cell biology is how cells cope with rapid changes in their environment. The range of molecular responses includes synthesis of transcription factors, changes in mRNA expression pattern which promote cell survival and, depending on the type of stress experienced by the cells – either increase or decline in protein synthesis.

In our project, we're focusing on two slightly different types of stress: hypoxia and accumulation of unfolded protein. We're using experimental datasets involving differential expression of mRNAs in hypoxia and unfolded protein response over a time course as input for computational models of signal transduction. Those models consist of a system of various numbers of non-linear ordinary differential equations describing temporal evolution of the key variables (presence and occupancy of transcription factor-binding motifs in the sequences of gene promoters, the TF target gene expression or external stimuli such as drugs used to induce stress) involved in either: regulation of protein synthesis, activation of hypoxia response-specific or unfolded protein response-specific genes.

Initial models and preliminary results will be presented.

BPscore: an effective metric for meaningful comparisons of structural chromosome segmentations

Rafał Zaborowski[1] and Bartek Wilczyński[1]

[1]*Institute of Informatics, University of Warsaw, Banacha 2, 02-967 Warsaw, Poland*

3D genome architecture studies gained significant attention in recent years. This is mainly due to rapid development of Chromosome Conformation Capture and its derivative methods like Hi-C, 4C and 5C . An interesting phenomena emerging in analysis of Hi-C data is occurrence of Topologically Associating Domains (TADs), which are regions of self enriched interactions indicating kilo- to megabase segmentation of chromosomes . Numerous studies in different cell types and species suggest TADs function as regulatory units preventing spread of epigenetic marks or distal regulatory elements therefore influencing the process of gene transcription.

As the number of such studies grows steadily and many of them aim to capture differences between TAD structures observed in different conditions, there is a growing need for good measures of similarity (or dissimilarity) between chromosome segmentations.

In this study we present a BP score, which is a relatively simple distance metric based on the bipartite matching between two segmentations. We provide the rationale behind choosing specifically this function and show its results on several different datasets, both simulated and experimental. Furthermore we show that the BP score is a proper metric satisfying the triangle inequality. Additionally, we analyze the local contributions to the BP metric and show that in actual comparisons between real datasets, the local BP score is correlating with the observed changes in the epigenetic marks.

Lieberman-Aiden E. et al.: Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326, 289 (2009)

Dixon J. R. et al.: Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376 (2012)

Nora, E. P. et al.: Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 7398 (2012)

Andrey, G. et al.: A switch between topological domains underlies HoxD genes collinearity in mouse limbs. Science 340, 6173 (2013)

Le Dily F. et al.: Distinct structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. 28, 2151 (2014)

Hepatic and pituitary gland gene expression profiling of candidate genes for metabolic disorders in Polish HF and Polish Red cattle

Dominika Wysocka[1] and Przemyslaw Sobiech[1]

[1]*Department and Clinic of Internal Diseases, Faculty of Veterinary Medicine,*
University of Warmia and Mazury in Olsztyn, 10-719 Olsztyn

Background: Metabolic disorder is a major health problem in dairy cattle, particularly to high milk producing dairy cattle. It is worthily emphasized that metabolic diseases have a very complex etiology and pathogenesis, and the impact of these diseases on hepatic and pituitary gland gene expression and organism oxidative balance is not fully described. Based on the RNA-seq experimental data, we first performed the screening of the candidate genes for metabolic disorders in bovine liver and pituitary gland tissues, followed by the comparison of hepatic and pituitary gland gene expression profiling of candidate genes in Polish HF and Polish Red cattle.

Methods: The study is aimed to determine the hepatic and pituitary gland expression of potential candidate genes which were participating in the maintenance of oxidative balance, negative nitrogen balance, as well as ketosis in Polish HF and Polish Red cattle. The RNA-seq experimental design comprised of young bulls aged between 6 to 12 months were investigated. For each breed, six liver and pituitary gland tissues were sequenced using Next-seq 500 illumina platform. The RNA-seq expression data were normalized by the reads per kilobase of exon per million reads mapped (RPKM) method. In context to metabolic disorder in cattle, we have investigated the following candidate genes to look the gene expression profiling in Polish HF and Polish Red cattle: SOD1, SOD2, SOD3, GPx2, GPx3, GPx5, GPx6, GPx7, GPx8, BDH1, FN1, ACSL3, HMGCL, HMGCS2, BDH2, ACSL6, ACAT2, IDH3B, ACAT1, HMGCS1, ACSL4, ACSL1, PC, CPT1A, OXCT1, and ACSL5 respectively.

Results: By comparing the RNA-seq data of metabolic tissues, study revealed that the investigated candidate genes were highly expressed in the hepatic tissues than to pituitary gland in cattle breeds. However, by comparing the Polish HF and Polish Red cattle, results revealed a similar trend of gene expression profiling of all investigated candidate genes for both metabolic tissues. Based on the obtained results, we have categorized gene expression profiling as: highly, moderately and averagely expressed candidate genes. In case of hepatic gene expression profiling, the SOD1, FN1, HMGCL, HMGCS2, ACAT2, ACAT1, HMGCS1, ACSL1 and ACSL5 were highly expressed (FPKM values of >a40), followed by SOD2, GPX3, IDH3B, PC and BDH2 as moderately expressed (FPKM values: >10 to <40), and averagely expressed SOD3, GPX5, GPX6, GPX7, GPX2, GPX8, BDH1, ACSL3, ACSL6, ACSL4, CPT1A, and OXCT1 respectively, in Polish HF and Polish Red breeds. In case of pituitary gland gene expression profiling, the SOD1 and GPx3 were highly expressed (FPKM values of >40), followed by SOD2, GPX8, IDH3B, ACAT1, ACSL4, and PC as moderately expressed (FPKM values: >10 to <40), and averagely expressed SOD3, GPX3,GPX5, GPX6, GPX7, GPX2, BDH1, BDH2, ACSL3, ACSL6, CPT1A, OXCT1, FN1, HMGCL, HMGCS2, ACAT2, ACAT1, HMGCS1, ACSL1 and ACSL5 respectively, in Polish HF and Polish Red breeds.

Concluding remarks: Based on this presented preliminary results on hepatic and pituitary gland gene expression profiling study, a further research plan is an essential pre-requisite to validate the Identified candidate genes. Furthermore, the understanding the genetic factors that predispose metabolic disorders in cattle would benefit the dairy industry as a whole by providing producers, breeding services, and veterinarians a tool to forecast a cow's susceptibility to metabolic disorders.

Assesing the impact of profiling methods in cancer studies

Agata Muszyńska,[1] Paweł Łabaj,[1,2,3] and Joanna Polańska[4]

[1] *Malopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland*
[2] *Austrian Academy of Sciences, Vienna, Austria*
[3] *Chair of Bioinformatics RG, Boku University, Vienna, Austria*
[4] *Data Mining Group, Institute of Automatic Control,*
Silesian University of Technology,Gliwice, Poland

The improvement of microarray calibration methods is an essential prerequisite for quantitative expression analysis. The aim of this project is to assess the impact of using different approaches on the final outcome. The program was written in R using multiple Bioconductor's packages. There are three stages of low- level analysis of gene expression data. Those stages are background correction, normalization and summarization. For each step two different methods were tested. High-level analysis aims in detecting significant biological changes between conditions. For this stage ANOVA technique was used. To adjust for multiple testing three correction methods were used: Holm, Benjamini- Yekutieli and Benjamini- Hochberg. In total 8 combinations of preprocessing methods and 3 multiple testing corrections were examined. To choose the best approach a list of obtained genes was compared to those reported in the previous articles. Method chosen as the best one was then used to validate the results based on Gene Ontology terms and KEGG pathways. In the end Hook_Quantile_PLM method with Benjamini- Hochberg correction was chosen. It detected 53 out of 71 previously reported genes and the outcome of functional analysis also confirmed information provided by reference articles.

SimRNP: a new method for fully flexible modeling of protein-RNA complexes and for simulations of RNA-protein binding

Michał Boniecki,[1] Nithin Chandran,[1] and Janusz Bujnicki[1,2]

[1] *International Institute of Molecular and Cell Biology in Warsaw, Poland*
[2] *Adam Mickiewicz University in Poznań, Poland*

Macromolecular complexes composed of proteins and nucleic acids play fundamental roles in many biological processes, such as the regulation of gene expression, RNA splicing and protein synthesis. Structures of some of these complexes have been experimentally determined, providing insight into mechanisms of their biological activities. However, for a great majority of protein-nucleic acid complexes, high-resolution structures are only available for some isolated components, often accompanied with low-resolution information about the overall shape (e.g. from cryo-EM or SAXS) or about the proximities and interactions of these components (e.g. from chemical cross-linking experiments). Given the scarcity of experimentally determined structures, computational techniques can be used to integrate heterogeneous pieces of information, guide structure elucidation and subsequently determine the mechanisms of action and interactions between the components.

Recently, we combined our approaches for protein and RNA modeling, and developed a method for modeling of proteins, RNAs, and protein-RNA complexes. SimRNP uses a coarse-grained representation of protein and RNA molecules, utilizes the Monte Carlo method to sample the conformational space, and relies on a statistical potential to describe the interactions in the folding process. It allows for modeling of complex formation for assemblies comprising two or multiple protein and RNA chains. The method allows for fully flexible modeling of protein-RNA binding, e.g. with components of unknown structure or which are disodrered in isolation. Modeling system can be supported by various type of restraints, that can be derived from biological experiments or just restrains that limit possible deformation of a given parts of the modeling system.

Nonadiabatic simulations of carbon monoxide photodissociation in H64Q neuroglobin

Jakub Rydzewski[1] and Wiesław Nowak[1]

[1] *Institute of Physics, Nicolaus Copernicus University, Poland*

Carbon monoxide (CO) is a leading cause of poisoning deaths worldwide, without available antidotal therapy. Recently, a potential treatment for CO poisoning was introduced, based on binding of CO by neuroglobin (Ngb) with a mutated distal histidine (H64Q). Here, we present an atomistic mechanism of CO trapping by H64Q Ngb revealed by nonadiabatic molecular dynamics. We focused on CO photodissociation and recombination of CO to wild type (WT) and H64Q Ngb. Our results demonstrate that the distribution of CO within the proteins differs substantially due to rearrangement of amino acids surrounding the distal heme pocket. This leads to the decrease of the distal pocket volume in H64Q Ngb in comparison to WT Ngb, trapping migrating CO molecules in the distal pocket. We show that the mutation implicates the shortening of the time scale of CO geminate recombination, making H64Q Ngb 2.7 times more frequent binder than WT Ngb.

Cytochrome P450 and their substrates: small molecule docking with Rosetta

Monika A. Kaczmarek,[1] Joanna M. Macnar,[2, 1] and Dominik Gront[1]

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

[2] *College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw*

P450 enzymes belong to the family of proteins involved in electron transport in a process called oxidative phosphorylation. They possess monooxygenase activity, thanks to which one atom of the oxygen molecule is introduced into substrate and the other to the water molecule. However, the presence of a second protein, an electron donor, is required for the reaction to take place. One of such proteins is ferredoxin. These enzymes catalyze the biosynthesis of endogenous compounds, mainly lipids. They metabolize and oxidize toxic compounds and xenobiotics representing about 75% of drugs. The catalytic center is hem.

The main object of our research is CYP109B1 belonging to the group of cytochromes P450 used in many biotechnological works due to the undiscovered electron transfer mechanism or the lack of accurate substrate binding analysis. In our research we used computer simulation methods to predict where the known ligands interact with CYP109B1. Thanks to the experimentally determined data, we can verify the predictions from the theoretical methods. The next step is to improve the in silico methods to better reproduce experimental data.

Docking Studies in Personalized Medicine: Photoactivation of the Anti-Diabetes Sulfonylurea Drug JB253 and Its Interactions with the Epac2 and SUR1 Proteins

Lukasz Peplowski,[1] Jakub Rydzewski,[1] Katarzyna Walczewska-Szewc,[1] and Wiesław Nowak[1]

[1]*Institute of Physics, Faculty of Physics, Astronomy and Informatics,*
Nicolaus Copernicus University, Torun, Poland

Photopharmacology is based on optical control of drugs. Drugs may be activated by an interaction with light leading to a conformational change. However, the physiological effect of the activated form may depend on individual features of the drug docking place[1,2]. In some cases, point mutations present in certain subject genome may render such drug to be ineffective. Thus, precise understanding of structural determinants of drug activity is very important for modeling, and structural bioinformatics.

In type II diabetes sulfonylurea drugs (SU) help to keep insulin level within an optimal range. It is rather difficult to synchronize the intake of SU with non-monotonous demand for insulin. New SU derivatives, as with JB253, are based on linking of a normal drug with the azobenzene light-sensitive chromophore, open new opportunities for the optical control of insulin release through trans–cis photoisomerization. JB253 is suspected to bind to a cytoplasmic protein Epac2 protein, possibly involved in the activation of insulin release[3]. The natural target of JB253 is sulfonylurea receptor protein SUR1, coupled with Kir6.2 voltage dependent potassium channel.

In this study, we aim to determine interactions of this photo-switch with the medically relevant proteins Epac2 and SUR1. The Hartree-Fock method was used to determine the CHARMM force field parameters of JB253. We performed docking of JB253 (in cis and trans conformations) to Epac2 and SUR1 to pinpoint possible allosteric effects related to the cis-trans isomerization. Short and preliminary molecular dynamics simulations of Epac2 with both isoforms docked were run and a 'sudden photoisomerization' model was applied to mimic the photoactivation.

Our data should help to optimize new light-activated drugs and to draw attention of the medical community to those fragments of EPAC2/SUR1 genes which may affect sulfonylurea drugs effectivity. Acknowledgements: This research was supported by the National Science Centre, Poland

(NCN) via 2016/23/B/ST4/01770 (WN) and 2015/19/N/ST3/02171 (JR).

[1] J. Broichhagen, et al. Nat Commun. 2014 Oct 14;5:5116.
[2] M.M Lerch , et al. Nat Commun. 2016 Jul 12;7:12054.
[3] Z. B. Mehta, et al. Sci Rep. 2017 Mar 22;7(1):291.