BOOK OF ABSTRACTS

XI Symposium of Polish Bioinformatics Society

September 5-7, 2018 - Wrocław, Poland

ptbi2018.pwr.edu.pl

Dear Colleagues

I have the great pleasure to welcome you all at the eleventh Symposium of the Polish Bioinformatics Society (PBS). This is a very special meeting since it marks the tenth anniversary of the first one. It was organised by a team from the Institute of Informatics at the University of Warsaw in Jadwisin near Warsaw in October 2008. The meeting was attended by almost 60 participants. It was organized in the formula that was a continuation of the informal PhD workshops previously organised by the teams of Jerzy Tiuryn (University of Warsaw) and Jacek Błażewicz (Poznań University of Technology). PhD students and MSc students as well as freshly minted Doctors presented orally their research, with ample time for discussion.

All subsequent Conventions, which changed the official name to Symposia starting from the fifth edition, were built on that tradition. This is a place where young bioinformatics practitioners can present their research on a forum broader than the colleagues from their own laboratory or institute.

Symposia have a great significance for the integration of the bioinformatics community in Poland. During our short history Symposia were organised in (or in the vicinity of) Warsaw (Jadwisin 2008), Poznan (Będlewo 2009), Gliwice (Ustroń 2010), Kraków (2011), Gdańsk (2012) and Wrocław (2013). In 2014, the Symposium returned to Warsaw as an integral part of the first BIO convention organised by four scientific communities - biochemists, biophysicists, cell biologists and bioinformaticians. In 2015 we organised the Symposium in Lublin, together with the Polish Society of Medical Chemistry. In 2016 the Symposium was held in Białystok and in 2017 in Uniejów (near Łódź, but organized by a team from from Poznań University of Technology. In 2018 the Symposium returned to Wroclaw, organised by a team led by prof. Małgorzata Kotulska from Wrocław University of Technology.

Our Symposia are changing and growing with us. In 2008 the community was very young. The absolute majority of participants in the Congresses were students, doctoral students and young doctors. But time does not stand still. From young doctors, young habilitated doctors were born, previous students are now doctors, who are not so young anymore. Symposia reflect this chance. Therefore, we have introduced opportunity to present for more experienced colleagues.

In the first two editions only oral presentations were admitted and prize for the best presentation was awarded. However, with the increasing number of participants it was impossible to give the opportunity to present for all willing participants. Thus, starting from the third edition we have added a poster session to the programme. This proved to be very successful scientific and social event awaited by all participant. It also gives us a chance for to award the best poster presenter.

The Symposium is also an opportunity to get acquainted with interesting achievements in related fields. To the event we invite their prominent representatives recruited from local scholars and our distinguished foreign collaborators.

As the most important annual events in the life of the Society, PBS Symposia are also used as a forum where we present awards to the winners of the best PhD and MSc theses in bioinformatics defended in the previous calendar year in Poland.

On behalf of the Board of PBS I would like to thank members of the Organising and Program Committees for their efforts. I hope that the eleventh edition of our Symposium will be the best one yet.

We will have the opportunity to listen to three invited lectures by our distinguished guests, four invited presentations by laureates of competitions for best PhD and MSc thesis in bioinformatics defended in 2017, twenty-one contributed talks and ten flash poster introductions. Finally forty posters will be displayed giving a lot of material for discussions during poster session.

To the presenters I wish successful presentations and excellent posters. To all participants I wish excellent atmosphere and fruitful scientific discussions during the entire conference.

Witold Rudnicki
President of the Board
Polish Bioinformatics Society

Contents

Program committee

Dr hab. Małgorzata Kotulska
Wroclaw University of Science and Technology (chair)

Dr hab. Tadeusz Andruniow
Wroclaw University of Science and Technology

Dr inż. Witold Dyrka
Wroclaw University of Science and Technology

Dr inż. Aleksandra Gruca
Silesian University of Technology

Dr hab. Paweł Mackiewicz
University of Wroclaw

Dr hab. Piotr Młynarz
prof. nadzw. Wroclaw University of Science and Technology

Prof. dr hab.Wiesław Nowak
Nicolaus Copernicus University in Torun

Dr hab. Witold Rudnicki
University of Bialystok & University of Warsaw

Prof. dr hab. Andrzej Sokalski
Wroclaw University of Science and Technology

Dr hab. Marta Szachniuk
Poznan University of Technology

Prof. dr hab. Joanna Szyda
Wroclaw University of Environmental and Life Sciences

Dr Bartosz Wilczyński
University of Warsaw

Organizing committee

Małgorzata Kotulska
Wroclaw University of Science and Technology (chair)

Witold Dyrka
Wroclaw University of Science and Technology

Przemysław Gagat
University of Wroclaw

Paweł Kędzierski
Wroclaw University of Science and Technology

Bogumił Konopka
Wroclaw University of Science and Technology

Barbara Kosińska-Selbi
Wroclaw University of Environmental and Life Sciences

Magda Mielczarek
Wroclaw University of Environmental and Life Sciences

SKNBIO INFORM A-TYKÓW UPWROC

The book of abstracts was created by Supervisor (Magda Mielczarek) & President (Bartosz Czech) of the Bioinformatics Student Association at Department of Genetics, Wroclaw University of Environmental and Life Sciences (Wroclaw, Poland)
http://theta.edu.pl/skn-bioinformatykow/

Partners & sponsors

Organizers

Polish Bioinformatics Society (PTBi)

Wroclaw University of Science and Technology

University of Wroclaw

Wroclaw University of Environmental and Life Sciences

Founding

KNOW consortium of Wroclaw Center for Biotechnology

Wroclaw Center for Biotechnology

Ministry of Science and Higher Education*

*Organization of the XI Symposium of the Polish Bioinformatics Society - a task financed under the agreement 939 / P-DUN / 2018 from the funds of the Minister of Science and Higher Education designated for dissemination activities.

Keynote speakers

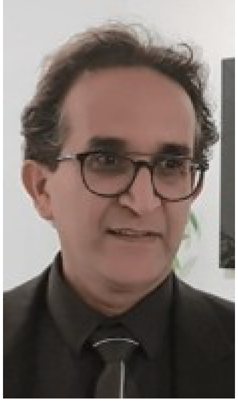
Johannes Söding

Johannes Söding obtained his PhD in 1996 in laser cooling of neutral atoms at the MaxPlanck Institute for Nuclear Physics and did postdoctoral experimental work on Bose-Einstein condensation of neutral atoms at the École Normale Supérieure in Paris. After three years as consultant at the Boston Consulting Group, he returned to science in 2002. He started his career in bioinformatics with Andrei Lupas at the Max Planck Institute for Developmental Biology in Tübingen, working on protein evolution, remote homology detection and structure prediction. In 2007 he became an independent research group leader at the Gene Center of the University of Munich (LMU). Since 2014 he leads the research group Quantitative and Computational Biology at the Max Planck Institute for Biophysical Chemistry. His group develops statistical and computational methods for analyzing data from high-throughput biological experiments, in particular for protein function and structure prediction, sequence search and assembly in metagenomics, transcription regulation, gene regulatory networks, and systems medicine.

Joanna Polańska

Joanna Polańska is a full professor and head of the data mining laboratory at Institute of Automatic Control of Silesian University of Technology. She obtained PhD in automatics and robotics from Silesian University of Technology in 1996, and habilitation in biocybernetics and biomedical engineering from Polish Academy of Sciences in 2008. Her research interests focus on development of bioinformatics algorithms and software tools for gene expression analysis and on computational intelligence methods for medical imaging. She is a member of the Polish Academy of Sciences and winner of the Silesian Scientific Award in 2017.

Mounir Tarek



Mounir Tarek is a Senior Research Director at the CNRS-Université de Lorraine, recipient of a Ph.D. in Physics from the University of Paris in 1994. He joined the CNRS after a four-years tenure at the National Institute of Standards and Technology (Gaithersburg Maryland USA) following three years tenure at the University of Pennsylvania). His research focus and expertise lies in the study of cell membranes transport processes. It involves the use of computational chemistry methods to study membranes, proteins, ion channels and membrane transport proteins. Over the last few years, he studied many aspects of electroporation of cell membranes subject to high electric fields. M. Tarek is a founding member and a member of the Scientific Council of the European Associated Laboratory EBAM ‚Pulsed Electric Fields Applications in Biology and Medicine'.

PRESENTATIONS

Keynote lecture B1

New algorithms and tools for large-scale sequence analysis of metagenomics data

Johannes Söding[1]

[1]Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Sequencing costs have dropped much faster than Moore's law in the past decade. The analysis of large metagenomic datasets and not their generation is now the main time and cost bottleneck. We present three methods that together allow us to move from an experiment-by experiment analysis to large-scale analyses of hundreds or thousands of metagenomic datasets. MMseqs2 [1] is a protein sequence and profile search method slightly more sensitive than PSI-BLAST and 400 times faster. MMseqs2 can annotate 1.1 billion sequences in 8.3 hours on 28 cores. MMseqs2 offers great potential to increase the fraction of annotatable (meta)genomic sequences. Linclust [2] is a sequence clustering method whose run time scales linearly with the input set size, not nearly quadratically as in conventional algorithms. It can cluster 1.6 billion metagenomic sequence fragments in 10 hours on a single server to 50% sequence identity, >1000 times faster than has been possible previously. PLASS (unpublished) is a metagenomic protein sequence assembler whose runtime and memory scale linearly with dataset size. It can assemble ten times more protein sequences from soil metagenomes, and faster than Megahit and other popular nucleotide-level assemblers. If time allows, I will also motivate our interest to contribute statistical methods to analysing large-scale medical data.

[1] Steinegger M and Soeding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, doi: 10.1038/nbt.3988 (2017)

[2] Steinegger M and Soeding J. Clustering huge protein sequence sets in linear time. biorxiv, doi: 10.1101/104034 (2018) (Nature Communications).

Session B1-1 Linguistic modeling in bioinformatics

Towards an Encyclopedia of Sequence Biology

Alexander Bolshoy[1]

[1]University of Haifa, Haifa, Israel

In this review I would present several topics relevant to a future of the scientific field that we propose to call it Sequence Biology. In some relevant to our topic works this field was called DNA Linguistics. At the heart of Sequence Biology lies a concept of a Sequence Code. In this review I would discuss three concepts: a concept of Sequence Biology, a concept of Encyclopedia of Genetic Codes, and a concept of Corpus DNA Linguistics.

Parasitism and peptide vocabularies

Michaela Zemková[1]

[1]Charles University in Prague, Prague, Czech Republic

Proteome (a set of all proteins of an organism) can be decomposed into a list of potential "words" (oligopeptides) of a given length. The list of all different oligopeptides represents actual peptide vocabulary of an organism. This linguistic-like approach is normally used in language processing, information retrieval systems etc. and was originally used by Shannon in 1948 as n-gram text decomposition. Here, we want to demonstrate the differences between peptide vocabulary usage of parasites and free-living organisms which can be driven by evolutionary arm-race between parasites and vertebrate hosts possessing MHC-based immunity system. Self-nonself discrimination in vertebrates is based on detection of peptides in proteins of parasites which are not present in proteins of a host. Therefore, parasitic organisms are under a strong selection pressure to eliminate maximum possible number of peptides-the potential targets for the hosts' immunity-from their proteomes. Using the concept of vocabulary usage, we tried to answer the question if parasites really modify their peptide vocabulary and we searched for indices of reduced peptide vocabulary in parasites by comparing the proteomes of 38 endoparasites with 33 free-living eukaryotic organisms and peptide length of 4 to 12 amino acids. We found that parasites really differ from free living organisms in diversity of 4-6 amino acids long peptides and use significantly impoverished pentapeptide vocabulary. This result is in accordance with the fact that the MHC I immunity recognition system works with strings of 4-5 amino acids. Besides practical impact of this results, we would like to point out that there is a bit problematic usage of such "linguistic" analogies in biology and that they have their advantages and limits.

Session B1-2

Monte Carlo Feature Selection and Interdependencies Discovery (MCFS-ID) with rmcfs 1.3.0

Michał Dramiński[1]

[1]Polish Academy of Sciences, Warsaw, Poland

MCFS-ID(Monte Carlo Feature Selection and Interdependency Discovery) is a Monte Carlo method-based tool for feature selection. It returns a ranked list of informative features, and thus play a significant role in the classification of objects that belong to different classes. This is achieved through constructing thousands of decision trees. MCFS-ID also allows for the discovery of interdependencies between the features, visualized as a directed graph of the pairwise interdependencies found. The discovered interdependencies thus provide a basis for making causal hypotheses to be verified using background knowledge. The effectiveness of the MCFS-ID approach in finding biologically/clinically relevant features and interactions between them has been confirmed in several studies. MCFS-ID is particularly suitable for the analysis of high-dimensional, 'small n large p' transactional and biological data. Its implementation rmcfs is publicly available on CRAN repository. The upcoming rmcfs version 1.3.0 implements: weighting of input attributes and a new highly efficient heuristic to evaluate millions of input attributes and provide the result in a reasonable amount of time. Weighting of attributes helps to discover some unique interdependencies between features that a user is highly interested in, e.g.,interdependencies between tens of patients' clinical features and hundreds of thousands of DNA methylation sites. The aim of the presentation is to introduce MCFS-ID algorithm and illustrate practical usage of the rmcfs package and its new features available in the upcoming version 1.3.0.

The Wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution

Szymon Majewski[1]**, Michał Ciach**[2]**, Michał Startek**[2]**, Wanda Niemyska**[2]**, Błażej Miasojedow**[1,2]**, Anna Gambin**[2]

[1]Polish Academy of Sciences, Warsaw, Poland

[2]University of Warsaw, Warsaw, Poland

We propose a new approach for the comparison of mass spectra using a metric known in the computer science under the name of Earth Mover's Distance and in mathematics as the Wasserstein distance. We argue that this approach allows for natural and robust solutions to various problems in the analysis of mass spectra. In particular, we show an application to the problem of deconvolution, in which we infer proportions of several overlapping isotopic envelopes of similar compounds. Combined with the previously proposed generator of isotopic envelopes, IsoSpec, our approach works for a wide range of masses and charges in the presence of several types of measurement inaccuracies. To reduce the computational complexity of the solution, we derive an effective implementation of the Interior Point Method as the optimization procedure. The software for mass spectral comparison and deconvolution based on Wasserstein distance is available at https://github.com/mciach/wassersteinms.

predPCR: automated classification of sigmoid curves

Michał Burdukiewicz[1], **Andrej-Nikolai Spiess**[2], **Stefan Rödiger**[3]

[1]University of Wroclaw, Wroclaw, Poland

[2]University Hospital Hamburg-Eppendorf, Hamburg, Germany

[3]Brandenburg University of Technology Cottbus-Senftenberg, Cottbus, Germany

Quantitative Real-Time PCR (qPCR) has a considerable popularity as a simple and robust method in research and precision medicine. The analysis of qPCR data is usually performed using well-defined methods, but the final interpretation of the result belongs to the experimentalist. However, with the advent of high-throughput qPCR devices, there is a growing need for the automated classification of samples. Herein, we describe predPCR, an ensemble classifier able to properly label qPCR curves. To create a universal tool, not biased to the single type of a qPCR device or set experimental conditions, we created a number of stratified datasets. During the demanding nested cross-validation procedure we identified areas of the competence of individual and found out which learning algorithms contribute to the most efficient ensemble model (AUC ~ 0.96). Our model was verified in vitro using qPCR runs from dilution experiments and later applied to the population-wide microRNA screening. predPCR is available as a web server http://www.smorfland.uni.wroc.pl/shiny/predPCR/. The local version is included in the PCRedux package (https://github.com/devSJR/PCRedux) for the open source statistical computing language and environment R.

TADeus — tool for clinical evaluation of chromosomal rearrangements modifying chromatin organization

Barbara Poszewiecka[1], **Paweł Stankiewicz**[2], **Rafał Płoski**[3], **Tomasz Gambin**[4], **Anna Gambin**[5]

[1] University of Warsaw, Warsaw, Poland

[2]Baylor College of Medicine, Houston, TX, US

[3] Medical University of Warsaw, Warsaw, Poland

[4]Warsaw University of Technology, Warsaw, Poland

Balanced chromosomal rearrangements with breakpoints in non-coding regions may convey an abnormal phenotype by position effects. Alterations in chromosome structure can disrupt interactions between gene promoters and their control elements or place genes in an inappropriate chromosomal domain. Clinical evaluation of the impact of such genomic aberrations on gene expression is often neglected, especially when other chromosomal rearrangements affecting genes are detected. Another reason for not considering such rearrangements is lack of appropriate methods and tools to visualize and prioritize neighboring genes according to their clinical relevance. To address these challenges, we developed TADeus, an easy-to-use open-source web application dedicated to clinical evaluation of genes whose expression may be affected by balanced chromosomal rearrangements. The main functionality of TADeus provides a view of regulatory landscape of regions surrounding breakpoints of the rearrangements by combing chromatin interaction data (Hi-C) with other omics data from different sources. TADeus prioritizes genes using haploinsufficiency data, number of broken enhancer-promoter connections and inclusion in the same Topologically Associating Domains (TADs). In addition TADeus can serve as a genomic browser that shows Hi-C matrices with one-dimensional genomic data and allows users to compare Hi-C data from different tissues.

PhyMet2: a database and toolkit for phylogenetic and metabolic analyses of methanogens

Jarosław Chilimoniuk[1], **Michał Burdukiewicz**[1], **Przemysław Gagat**[1], **Sławomir Jabłoński**[1], **Michał Gaworski**[1], **Paweł Mackiewicz**[1], **Marcin Łukaszewicz**[1]

[1]University of Wroclaw, Wroclaw, Poland

Background: Innovations in DNA sequencing technologies allowed for the rapid development of metagenomics, DNA sequencing of environmental samples, and consequently, identification of a plethora of new uncultivated microorganisms. However, in order to describe the phenotype of microorganisms, i.e. to gather data on their physiology, morphology, and biochemistry, the metagenome analyses need to be supplemented with studies of microorganisms isolated in pure culture. Unfortunately, searching for the optimal culturing conditions is expensive, time-consuming, and technically difficult. An important component of the understudied ecosystems are methanogens, archaea producing methane, a potentgreenhouse-effect gas. Therefore, we created PhyMet2,the first database that combines descriptions of methanogens and their culturing conditions with genetic information. Methods: In order to train MethanoGram, we used n-grams, i.e. subsequences of the length n that were extracted from 16S rRNA and mcrAnucleotide sequences deposited in the PhyMet 2 database. To estimate theculturing conditions we chose the random forests algorithm. Results: The database, with auser-friendly interface design, contains a set of utilities for interactivedata browsing and comparing as well as exploring phylogeny and searching forsequence homologues. The data contained in PhyMet2 was used todevelop a web server, MethanoGram that quickly and accurately predicts theconditions for optimal growth of methanogens: temperature, pH, and NaClconcentration, i.e. the key factors that shape the composition of methanogenic communities. Availability: PhyMet2with MethanoGram predictor is available at http://metanogen.biotech.uni.wroc.pl.

Poster: P31

Keynote lecture B2

Exploring the Complex Dynamics of an Ion Channel Voltage Sensor Domain via Computation

Mounir Tarek[1]

[1]CNRS, University of Lorraine, Nancy, France

Voltage sensors (VSDs) are ubiquitous domains of voltage-gated ion channels that act as transducers of transmembrane electric signals within and across excitable cells. As such they are implicated in key physiological processes e.g. cellular contraction, cardiac and neuronal electrical activity. Among the body of discoveries that have contributed to increase our knowledge of VSDs function was the recording of transient currents of very low amplitude called today "gating currents". These currents initially measured in the 1970s follow from the reorganization of charged residues of the protein in response to a change in the transmembrane voltage. Interpreting gating currents in light of simplified kinetic models has been for decades, and remains to this day, the main approach to investigate the molecular determinants of VSD activation in order to better understand voltage-gated ion channel function, as well as dysfunction that might result for instance from genetic mutations.

Here, we developed and alternative strategy in which we extracted from molecular models of the VSDs of the Kv1.2 Potassium voltage gate channel, the potential of mean force (PMF) that describes the energetics of their activation mechanism. We then deduce gating currents that are directly comparable to experimental recordings, showing how the Molecular Dynamics based kinetic models of VSD activation represents an innovative tool to answer questions regarding voltage-gated channel function. Our calculation of the voltage sensor PMF represents therefore a crucial milestone toward a quantitative, molecular-level picture of an ion channel gating.

Session B2-1

Investigating the dynamics of ATP-sensitive potassium channel gating: a molecular dynamics study

Katarzyna Walczewska-Szewc[1], **Wiesław Nowak**[1]

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

Despite being the main target of type 2 diabetes therapy, ATP-sensitive potassium channels (KATP) are yet unexplored. Their role in insulin secretion from pancreatic beta-cells is well-estabilished. Nevertheless, the exact mechanism of their action is unclear. The KATP assembly is composed of four inwardly rectifying K+ channel subunits (Kir6.2) and four sulfonylurea receptor (SUR1) moieties. The recent publications (2017) of structures of KATP at near-atomic resolution uncovered many molecular details regarding to the complex assembly. Moreover, the positions of several nucleotide binding sites (including sulfonulurea, a drug used in type 2 diabetes treatment) have been confirmed, which allows us to hypothesize about their role in the channel gating. The complexity of KATP system reinforces the need to use complex methods, like molecular dynamics (MD), to understand its action. We would like to show how different MD methods can be used to provide an insight into the spatial structure and real-time dynamics of system like that. Our studies can move us one step further into understanding the exact mechanisms of KATP channel gating.

Backbone Brackets and Arginine Tweezers delineate Class I and Class II aminoacyl-tRNA synthetases

Sebastian Bittrich[1], Florian Kaiser[2], Sebastian Salentin[2], Christoph Leberecht[1], V. Joachim Haupt[2], Sarah Krautwurst[1], Michael Schroeder[2], Dirk Labudde[1]

[1]University of Applied Sciences, Mittweida, Germany

[2]Biotechnology Center (BIOTEC), TU Dresden, Germany

Aminoacyl-tRNA synthetases (aaRS) are primordial enzymes essential for interpretation and transfer of genetic information. Understanding the origin of the peculiarities observed with aaRS can explain what constituted the earliest life forms and how the genetic code was established. These enzymes are responsible for loading tRNA molecules with the correct amino acid, required for translation during protein biosynthesis. Two Classes of aaRS can be observed today, whereby each is responsible for ten amino acids. A delicate balance between these Classes is manifested in two structural motifs: The Backbone Brackets and the Arginine Tweezers. Furthermore, aaRS constitute a reflexive system, whereby each Class cannot exist without its counterpart. The characteristics of both motifs support the Rodin-Ohno Hypothesis, which states that prototypic aaRS were once coded on complementary strands of the same gene. Both Classes differ significantly at the sequence and structure level, feature different reaction mechanisms, and occur in diverse oligomerization states. The one unifying aspect of both Classes is their function of binding ATP. We identified Backbone Brackets and Arginine Tweezers as most compact ATP binding motifs characteristic for each Class.

Geometric analysis shows a structural rearrangement of the Backbone Brackets upon ATP binding, indicating a general mechanism of all Class I structures. We demonstrate that only the combination of sequence, structure, and ligand binding information allowed us to identify these two motifs. This large-scale study of nearly 1000 protein chains demonstrates a possibility to evolutionary insights by employing the diverse toolkit of structural bioinformatics.

Evaluation of residue-residue contact prediction in metamorphic proteins

Mateusz Skrzypecki[1], Paweł Woźniak[1], Witold Dyrka[1]

[1]Wroclaw University of Science and Technology, Wroclaw, Poland

Metamorphic proteins are challenging for tools that predict residue-residue contacts. Such proteins have more than one tertiary structure for one amino acid chain. In contrast, the state-of-the-art contact prediction tools based on the direct-coupling analysis (DCA) try to predict one specific structure for one sequence. To investigate the issue in depth, we conducted qualitative analysis of pairs of proteins that had the same sequence, but different conformations. We noticed that predicted contacts from alternative conformations were mixed up or were indicated for only one conformation. therefore, based only on DCA results, contacts cannot be assigned to the specific tertiary structure. However, some insights can be deduced when analyzing DCA results for the arrangement of secondary structures relative to each other. Suppose we have a metamorphic protein that has two tertiary structures. They differ in that in one conformation two secondary structures are far apart, but in the second one, they interact with each other. With the current DCA effectiveness, it is possible to indicate with a high probability which secondary structures are mutually correlated. Having only the first conformation solved experimentally and comparing it with the results from DCA, we will see contacts predicted between secondary structures that are far from each other. This suggests the presence of an alternative conformation. Further, the pattern of predicted contacts between secondary structures may be used to find structural templates for the alternative conformation among proteins with solved conformations, and to select them for further analysis.

Keynote lecture B3

Single-cell sequencing — new data, new challenge

Michał Marczyk[1,2], Christos Hatzis[2], **Joanna Polańska**[1]

[1]Silesian University of Technology, Gliwice, Poland

[2]Yale Cancer Centre, Yale University, New Haven, USA

Single-cell sequencing is an important technology to define intercellular heterogeneity, rare cell types, cell genealogies, somatic mosaicism, microbes, and disease evolution. Initially, the method was applied by the Surani laboratory in 2009. Now, high-throughput technologies enable the profiling of hundreds of thousands of cells in parallel, allowing high-resolution analysis of individual cells and providing knowledge on the molecular background of biological processes that occur under the observed conditions. Single-cell cancer transcriptomics and genomics highly improve experimental sensitivity to map clonal evolution, track the development of therapy resistance, and analyse rare tumour cell populations such as tumour stem cells and circulating tumour cells.

A major advantage of scRNA-Seq is the unbiased identification of cellular subpopulations from heterogeneous populations of cells. The structure of a complex tissue is tightly linked with its function, so determining the frequency and identity of cell types is crucial. Additionally, single-cell transcriptomic analysis of transitions between cellular states (for example during development or differentiation) can reveal new insights into regulatory mechanisms. Transitions between states can be binary or gradual and can involve one or even multiple intermediate states. Understanding the nature of the process and possible intermediate states can lead to the identification of key genes that act as switches and drivers of these processes. Hence, clustering cells into groups by their gene expression levels is an important challenge that must be addressed with high priority in the computational pipelines. The aim of the work is evaluating existing methods and present new approaches for clustering high-dimensional single cell sequencing data.

A challenge in cancer research is to develop combination therapies, which gives the highest efficiency with reduced toxicity and no resistance to therapy. The long duration of treatment in vitro can give vital information on the sensitivity of cell lines, development of resistant clones and effect of sequencing and doses on pharmacodynamics. We focus on the analysis of triple negative breast cancer (TNBC) data on the single-cell transcriptomic level. TNBC is a highly aggressive and heterogeneous disease. The heterogeneity of these tumours implies different chemosensitivity to standard therapies, so the discovery of new, more effective therapies for all TNBC patients is indispensable. The presence of multiple different subclones may limit the response to targeted therapies and contribute to the acquisition of drug resistance. Application of single-cell RNA sequencing technology allows to evaluate heterogeneity between resistant clones and assess transcriptional similarities and differences within a population of cells.

Session B2-2 Flash

Drug activity assesment with metabolomics

Mariusz Bromke[1], **Jerzy Wiśniewski**[1], **Izabela Szczuka**[1], **Berenika Szcześniak-Sięga**[1], **Małgorzata Krzystek-Korpacka**[1]

[1]Medical University of Wroclaw, Wroclaw, Poland

One of ways of a new drug discovery is the chemical substitution modification of already existing and active drugs. In such an approach, the action of known reference drug and its derivatives must be thoroughly studied and analysed invitro before the best performing entities will go into further steps of pre-clinical studies. In presented here research project we have compared the action of piroxicam, a non-steroid anti-inflammatory, analgesic, and antipyretic drug of the oxicam group; with action of 5 chemically modified derivatives. The action of the drug relies mainly on inhibition of cyclooxygenases, which are responsible for synthesis of prostaglandins at the inflammation site. The activity of piroxicam and its derivatives was studied in cultures of CaCo-2 (human colorectal cancer) cells. For this purpose we have studied the perturbed metabolism with UHPLC-MS. We have applied two metabolomic data analysis approaches: targeted andnon-targeted. In the first one, we have focused on amino acids which in the core of a cell's metabolism. In the non-targeted approach, we have applied mathematical analysis (PCA, hierarchical clustering) in order to study and compare „metabolic finger prints" of the response to the treatment with piroxicam derivatives. Within few hours of treatment, we have observed a massive increase of asparagine and aspartic acid. Our analysis revealed also other unknown changing, yet significant, analytes, of which identity discovery is our next goal. The combination of metabolomics supported with mathematical tools will greatly complement or guide next targeted transcriptome analysis. The metabolomic data analysis was performed with use of scripts written in R, which among others utilised pca methods package from the Bioconductor.

Poster: P01

ModeLang — an approach to model dynamical systems in bioinformatics

Tomasz Prejzendanc[1], Szymon Wąsik[1], Jacek Błażewicz[1]

[1]Poznan University of Technology, Poznan, Poland

ModeLang is Controlled Natural Language that has been designed for modelling of dynamical systems in bioinformatics. Main way of modelling is based on systems of differential equations. Analysis of the models in large systems of differential equations can be inconvenient and meaning of the model can be unclear based on large set of parameters and variables. Controlled Natural Language can be the way to make the model self-documented and more approachable for new scientists. There are several examples of usage of ModeLang and its comparison to the systems of differential equations.

Poster: P11

New insights in Y chromosome degeneration

Dorota Mackiewicz[1], Piotr Posacki[1], Michał Burdukiewicz[1], Paweł Błażej[1]

[1]University of Wroclaw, Wroclaw, Poland

In the genetic system of sex determination, most mammalian females possess two copies of X chromosome, whereas males have one X and one Y chromosome. A spectacular phenomenon related to the sex chromosome evolution is the shrinkage of the Y chromosome. It was proposed that this process is related with the cessation of recombination between the Y and its counterpart, i.e. the X. The suppression of recombination occurred by the series of large-scale inversions happened most likely on the Y chromosome. Then selection favoured successive mutations and stepwise extension of the genetic linkage because it increased the probability of joint transmission of genes beneficial for one sex. To study all these aspects, we applied a more general and advanced computer simulation model in which the recombination rate between the sex chromosomes can freely evolve and individuals can create unfaithful or faithful pairs. We found that only under the unfaithfulness of mates the number of females increases at the expense of males inthe evolving population and the accumulation of mutations on the Y chromosome occurs. Therefore, the recombination rate between the X and Y decreases very quickly and the Y degenerates. Thus, the X chromosomes are cleaned off defective alleles and the reproduction potential of population, measured by the number of females, is not reduced. The simulation showed that the suppression of recombination is spontaneous and does not require inversions.

Poster: P17

Improving detection of gene duplications in whole-genome sequencing data using allelic depth imbalance

Paweł Sztromwasser[1], Tomasz Stokowy[2], Stefan Johansson[2], Vidar M. Steen[2], Kjell Petersen[2], Inge Jonassen[2]

[1]Medical University of Lodz, Lodz, Poland

[2]University of Bergen, Bergen, Norway

Copy-number variants (CNVs) are responsible for an equal share of genetic variation among humans as single-nucleotide variants (SNVs). However, detection and genotyping of CNVs using short-read sequencing is more challenging compared to calling SNVs. Four types of information are typically used to discover CNVs in whole-genome sequencing (WGS)data: read depth, read-pair orientation, split-reads, and loss of heterozygozity (in case of deletions). Bioinformatics tools utilize these types of information, alone or in combinations, but they provide variable performance and show low concordance between the results, especially for duplications. Here we explore the utilization of an additional type of information — allelic depth imbalance (ADI)- to improve detection of duplications. We used WGS data from a widely studied CEPH trio, and two complementary datasets of known duplications as gold-standards. The known duplications and a set of randomly selected negative control regions were ranked using the ADI score. The score was based on allele ratio in heterozygous SNV calls within the duplications and in control regions. Our results show that ADI score differs between diploid and higher copy-number state regions. The score was able to classify known duplications with 67% sensitivity and 6% precision. When combined with read-depth, ADI score achieved up to 9% improvement insensitivity and up to 7% decrease in false-positive calls, compared with read-depth alone. Although ADI alone cannot discriminate false-positives sufficiently, used together with other signals, it can improve detection of duplications in whole-genome sequencing data

Funding: National Science Centre, Poland (grant nr 2016/23/P/NZ2/04251), Bergen Research Foundation. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 665778.

Poster: P08

The standard genetic code robustness to both point and frameshift mutations

Małgorzata Wnętrzak[1], **Paweł Błażej**[1], **Paweł Mackiewicz**[1]

[1]University of Wroclaw, Wroclaw, Poland

Several hypotheses have been formulated to explain the origin and evolution of the standard genetic code. One of them states that the standard code evolution could have been driven by the need to reduce the harmful effects of amino acid replacements in proteins, caused by mutations in protein-coding genes or errors during their translation. However, most scientist studying this hypothesis focus solely on the robustness of the genetic code to single nucleotide substitutions, although more harmful consequences in the coded proteins result from deletions and insertions. Therefore, in our research we decided to check the combined influence of both point and frameshift mutations on the robustness of the standard genetic code. Using a genetic algorithm and multi-objective optimization we searched for codes optimized for minimization and maximization of errors caused by both types of mutation, regarding different amino acid properties. The features of the optimized codes were then compared with the properties of the standard genetic code, in order to assess its level of optimality in the whole space of theoretical codes. The results indicate the presence of a tendency to minimize the costs of amino acids replacements in proteins in the standard genetic code even though it is not fully optimized regarding this property. These results indicate that other factor must have influenced the assignments of amino acids to codons.

This work was supported by the National Science Centre, Poland (Narodowe Centrum Nauki, Polska) under the Grant number 2017/27/N/NZ2/00403.

Poster: P43

Bioinformatic analyses of mitochondrial genomes show duplicated regions in many parrot taxa

Aleksandra Kroczak[1], Adam Dawid Urantówka[2], Paweł Mackiewicz[1]

[1]University of Wroclaw, Wroclaw, Poland

[2]Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

It is commonly assumed that mitochondrial genomes of vertebrates evolve towards the reduction of their size or have already reached a steady state. However, recent reports about the mitogenomes of birds indicate that they were subjected to frequent duplications and rearrangements. Previous analyses showed that duplications of control region and adjacent genes are present only in several lineages of birds, in which these duplications occurred independently. However, our molecular and bioinformatic analyses contradict this view. We mapped the presence/absence of duplications in the mitogenomes onto parrot phylogeny and performed sensitive homology searches for potential pseudogenes in these regions. The results showed that the duplicated region was already present in early diverged groups of parrots. This state was inherited by the ancestors of main parrot lineages and was lost in later diverged groups. We also found that the duplicated control regions were subject to concerted evolution and their neighboring genes were pseudogenized or finally lost. Our research suggests that the maintenance of duplicated control regions can give a selective advantage due to more efficient initiation of replication or transcription and larger number of replicating genomes per organelle. It may allow a more effective energy production by mitochondria.

This work was supported by the National Science Centre Poland (Narodowe Centrum Nauki, Polska) under Grant no. 2015/17/B/NZ8/02402. Some computations were carried out at the Wroclaw Center for Networking and Supercomputing under the grant no. 307.

Poster: P16

Detecting Life Signatures with RNA Sequence Similarity Measures

Mateusz Kudla[1], Szymon Wasik[2], Natalia Szostak[2], Michal Wachowiak[1], Krzysztof Krawiec[1], Jacek Blazewicz[2]

[1]Poznan University of Technology, Poznan, Poland

[2]European Centre for Bioinformatics and Genomics, Poznan, Poland

Nowadays, the RNA World is the most plausible hypothesis for explaining the origins of life on Earth. The supporting body of evidence is growing and it comes from multiple areas. The latter often assume the existence of hypothetical species on the prebiotic Earth, and study the characteristics of a population of 'agents' representing such species, especially their replicative capabilities. However, it is often hard to verify whether or not a genetic sequence representing an agent has the characteristics of biological sequences, and is thus likely to be functional. The primary objective of the presented research was to verify the possibility of building a computational 'life probe' for determining whether a given genetic sequence is biological, and assessing the sensitivity of such probes to the signatures of life present in known biological sequences. Given the fundamental importance of information theory and its many successful applications in biology, we have proposed a decision algorithm based on the normalized compression distance and assessed its discriminative capabilities in comparison to the Levenshtein distance. We have validated the proposed method in the context of the RNA World hypothesis using short genetic sequences and demonstrated that both measures can be successfully used to construct life probes that are significantly better than random classifiers, while varying from each other when it comes to detailed characteristics. We also observed that fragments of sequences related to replication have better discriminatory power than sequences having other molecular functions. In a broader context, this work brings another piece of evidence that NCD is an interesting alternative to the existing means of comparing biological sequences.

Poster: P04

Bioinformatic study of RNA multi-branched loops

Jakub Wiedemann[1], Maciej Miłostan[1,2], Marta Szachniuk[1,2], Ryszard Adamiak[1,2]

[1]Poznan University of Technology, Poznan, Poland

[2]Polish Academy of Sciences, Poznan, Poland

In the last decade, we observe a number of new methods to predict three-dimensional structures of proteins and RNAs. With these methods, scientific society tried to address the problem of the disproportion between known sequences and structures. However, the prediction of 3D structures of biomolecules still needs a lot of improvement since its results are far from perfect. Existing 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 modelling 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].

Poster: P09

Analysis of gplmDCA tool and created filters performance for contact site prediction in knotted and metamorphic proteins

Julia Pelc[1], Paweł Woźniak[1], Mateusz Skrzypecki[1], Małgorzata Kotulska[1], Gerrit Vriend[2]

[1]Wroclaw University of Science and Technology, Wroclaw, Poland

[2]RadBoud UMC Nijmegen, Nijmegen, Netherlands

More and more attention is devoted to the development of *in silico* methods that will allow us to understand the tertiary structure of proteins based on the prediction of their contact sites. Contact sites are pairs of amino acids responsible for maintaining the spatial structure of proteins that are separated form each other by at least 4 amino acids in the sequence, but distant by at most 8 Angstroms in space. Very promising tools are those using the correlated mutation method in Multiple Sequence Alignment (MSA) in conjunction with Direct-Coupling Analysis (DCA). These methods point each pair in the protein - the higher the score is, the more likely is that the pair creates a contact site. In order to improve the accuracy of one of the DCA tools (gplmDCA) together with the team we have developed a set of filters that, based on available information about amino acids creating a pair, allow us to remove from the set of highest-rated pairs those that cannot create contact sites. In my work, I have examined the accuracy of both, gplmDCA and six of created filters on two groups: knotted and metamorphic proteins. For the gplmDCA tool the accuracy of correctly predicted contact sites amounted to 24% for tool test-set while for the knotted proteins it was 65% and for metamorphic ones - 8%. The best results of filters accuracy were obtained for intra-secondary structure and buried-exposed filters, accounting for respectively 73% and 80% for knotted proteins and 98% and 91% for metamorphic ones.

Poster: P27

G-quadruplex Topology from Bioinformatics Perspective

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

[1]Polish Academy of Sciences, Poznan, Poland

[2]Poznan University of Technology, Poznan, Poland

One of the current interest in biological and bioinformatic fields are non-canonical G-quadruplex structures. These formations exist in guanosine-rich DNA, RNA and nucleic acids analogs. Basic building unit of a G-quadruplex is G-tetrad, created by four guanines located in a pseudo-planar architecture. At least two G-tetrads stacked upon each other are necessary to compose the simplest G-quadruplex structure. 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, or hybrid-type. G-quadruplexes are 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. They are proved in vivo and in vitro, to be suitable targets especially in cancerous diseases. In our research, we analyzed structures, stored in PDB database, that are classified as the ones containing G-quadruplex formation within them. We proposed a new manner to visualize G-quadruplex secondary structures, and moreover, we prepared a new classification of G-quadruplexes. Now we are focusing on an algorithm for automatic classification of G-quadruplexes within the proposed representations. We visualized these structures and defined their strands orientation in G-quadruplexes by using the newest version of RNApdbee program and based on these results, we created 3D schemes of G-quadruplexes classes.

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

Poster: P39

Session B2-3

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

Juan Francisco Carrascoza Mayen[1], **Jakub Rydzewski**[2], **Natalia Szostak**[1,3], **Jacek Blazewicz**[1,3], **Wiesław Nowak**[2]

[1]Poznan University of Technology, Poznan, Poland

[2]Nicolaus Copernicus University in Torun, Torun, Poland

[3]IBCH, Polish Academy of Sciences, Poznan, 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 the simple precursors of nucleic acid bases or aminoacids, we performed quantum Car–Parrinello molecular dynamics simulations. We prepared four condensed-phase systems modeled as prototypes of the primordial soup. Montmorillonite clay (MMT) was used as a possible catalyst. We monitored the chemical reactions in both gas-like and MMT-confined simulation boxes on a 20-ps timescale at 1 atm and 300 K, 400 K, and600 K conditions. Elevated temperatures did not considerably affect the reactivity of the elementary components of the gas-like boxes; however, the presence of MMT substantially increased the formation probability of new molecules. An analysis of the atom—atom radial distribution functions indicated that the presence of Ca2+ ions at the surface of the internal cavities may be an important factor in the initial steps of the complex molecule formation at the early stages of the Earth's history and in those exoplanetary regions of the Universe where MMT materials are available.

The structure of the genetic code as an optimal graph clustering problem

Paweł Błażej[1], Dariusz R. Kowalski[2], Dorota Mackiewicz[1], Małgorzata Wnętrzak[1], Daniyah A. Aloqalaa[2], Paweł Mackiewicz[1]

[1]University of Wroclaw, Wroclaw, Poland

[2]University of Liverpool, Liverpool, United Kingdom

The specific structure of the standard genetic code is still a puzzle for biologists. In order to evaluate the quality of the genetic code structure, we introduce a new general methodology which comes from the graph theory. It allow us to analyse new properties of the genetic code. Especially, we describe new measures of the codongroup's robustness against changes in protein-coding sequences caused by single nucleotide substitutions. Following this approach the standard genetic code can be represented as a partition of an undirected and unweighted graph. Moreover, we present and discuss the structure of optimal genetic codes in terms of conductance, which measures the proportion of non-synonymous substitution to all substitutions. We show that the finding the optimal code corresponds to the solution of optimal graph clustering problem. Despite the fact that the standard genetic code is far from being optimal, its structure is characterized by many codon groups reaching desired properties. The standard genetic code represents most likely a local minimum in terms of errors occurring in protein-coding sequences and their translation.

This work was supported by the National Science Centre, Poland (Narodowe Centrum Nauki, Polska) under Grant number 2017/27/N/NZ2/00403 and 2017/01/X/NZ2/00608.

GBSC: Method for clustering proteins by repeats in sequences

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

[1]Silesian University of Technology, Gliwice, Poland

[2]University of Warsaw, Warsaw, Poland

[3]Polish Academy of Sciences, Warsaw, Poland

A significant portion of protein sequences is built out of repeats. Repeats are mainly found in Low Complexity Regions (LCRs). Such regions occur in about 14% of proteins [1]. Each repeat can be composed of single amino acid as well as multiple of them. Changes in number of repeats in protein sequence may affect protein function e.g. it can cause neural disorders [2]. Current methods for searching for similar proteins mask LCRs in order to improve align of homologous proteins. There is couple of methods known from literature which extract repeats from proteins. However, none of them is able to cluster such proteins by similar repeats. In order to overcome this limitation, we propose a new method for clustering sequences by repeats. The graph based on sequence clustering (GBSC) method is a new approach to analyse repeats in proteins. The proposed algorithm builds graphs from a sequence and create cycles from them. Nodes in graph are created from k-mers of the sequence and transitions connect neighboring k-mers. The positions of each related k-mers are tagged. Each node and transition which occurs more often than a given threshold constructs particular cycle. One sequence can be composed of more than one type of repeats, so in this case the algorithm constructs adequate number of cycles. At the end the method groups sequences which contain corresponding type of cycle for further investigation.

References:

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

2. Alexander E. Conicella, Gül H. Zerze, Jeetain Mittal, and Nicolas L. Fawzi. ALS mutations disrupt phase separation mediated by -helical structure in the TDP-43 low complexity C-terminal domain. Structure, 2016.

Aortic aneurysm formation and progression analyzed using Petri net-based approach

Kaja Chmielewska[1], **Dorota Formanowicz**[2], **Piotr Formanowicz**[1,3]

[1]Poznan University of Technology, Poznan, Poland

[2]Poznan University of Medical Sciences, Poznan, Poland

[3]Polish Academy of Sciences, Poznan, Poland

Despite formation of the abdominal aortic aneurysms (AAA) is relatively common vascular disorder, there is limited knowledge about the mechanisms responsible for this pathology. Etiology of AAA is multifactorial but commonly involves a weakening of the arterial wall, usually by atherosclerosis. This issue is very important and it seems that only taking into account aneurysms the complexity of this disease we can help in predicting the dynamics of this disease and risk of aneurysms rupture. Because the issue is complex, a systems approach has been used to build and analyze a Petri net-based model of the studied phenomenon. This approach allows to systematize existing knowledge about it and gives a better insight into this complex process. The analysis of models expressed in the language of Petri nets theory can be based on t-invariants. They correspond to sets of transitions (called supports of t-invariants) being counterparts of subprocesses which do not change the state of the modeled system. Searching for similarities among such invariants may lead to discoveries of some unknown properties of the analyzed biological system. In this work a method for a more detailed analysis of t-invarinats based on looking for subset of transitions of some minimal cardinality, which is included in sufficiently large number of t-invariants has been used to identify important subprocesses.

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

Parallel RNA World Simulator

Jarosław Synak[1], Natalia Szóstak[1], Sebastien Varrette[2], Szymon Wąsik[1], Pascal Bouvry[2], Jacek Błażewicz[1]

[1]Poznan University of Technology, Poznan, Poland

[2]University of Luxembourg, Luxembourg

Earth is a planet full of life and it has been like that since the beginning of the human history. Despite of that, the question about the beginnings of life still remains without the answer. Even the simplest cells possess very complicated machinery, which keeps them alive. How it could emerge spontaneously remains unclear. However, there are theories which try to explain the whole process, one of them is RNA World Hypothesis. According to this hypothesis, the whole life started with a population of RNA chains (formed spontaneously in a prebiotic soup). Proving or denying this hypothesis is extremely difficult, because the amount of time that passed is so long, that no direct evidence exists. One of the possibilities to research this subject are simulations. The authors propose a parallel, multiagent approach, based on Replicate-Parasite Model. The main goal is to check, under which circumstances the simulated population can survive, as it can be an important contribution into RNA World Hypothesis research. One of the main advantages of the presented approach is its credibility. Multiagent systems can represent molecular systems more accurately than the Cellular Automata. Another important advantage is the scalability, which allows to make more computations than in previous, sequential version, given the sufficient number of computational cores.

Session B3-1 Best PhD & MSc thesis awards

Modelling of protein dissociation in mass spectrometry

Michał Ciach[1]

[1]University of Warsaw, Warsaw, Poland

Mass spectrometry is a popular analytic technique of identification of chemical compounds. In proteomics, it is often used to identify proteins. Using a suitable technique of protein fragmentation allows to obtain much more information about its sequence. One of such techniques is the Electron Transfer Dissociation (ETD). Since the ETD is a relatively new technique, many questions about it still remain unanswered. They regard, among others, the precise mechanism of the reaction, the possibility of probing the structure, or the planning of experiments.

In this work we construct a formal mathematical model of the ETD and some of the co-occuring side reactions. The model allows quantitative study of this fragmentation technique. It is based on a stochastic description of the reaction for a single ion. We employ a populational approach to obtain a deterministic model of reaction for a big population of ions. Numerical optimization procedures allow to fit such model to experimental data processed by the MassTodon software. We show how to apply the model to find the experimental settings which give optimal fragmentation and minimize the side reactions. The model has been implemented in a program called ETDetective. The software has been made publicly availiable under the BSD 2-clause license.

Systematic classification of the His-Me finger superfamily

Jagoda Jabłońska[1]

[1]University of Warsaw, Warsaw, Poland

The His-Me finger endonucleases, also known as HNH or ββα-metal endonucleases, form a large and diverse protein superfamily. The His-Me finger domain can be found in proteins that play an essential role in cells, including genome maintenance, intron homing, host defense and target offense. Its overall structural compactness and nonspecificity make it a perfectly-tailored pathogenic module that participates on both sides of inter- and intra-organismal competition. An extremely low sequence similarity across the superfamily makes it difficult to identify and classify new His-Me fingers. Using state-of-the-art distant homology detection methods, we provide an updated and systematic classification of His-Me finger proteins. In this work, we identified over

100 000 proteins and clustered them into 38 groups, of which three groups are new and cannot be found in any existing public domain database of protein families. Based on an analysis of sequences, structures, domain architectures, and genomic contexts, we provide a careful functional annotation of the poorly characterized members of this superfamily. Our results may inspire further experimental investigations that should address the predicted activity and clarify the potential substrates, to provide more detailed insights into the fundamental biological roles of these proteins.

A challenge of automatic classification organic reactions yield and duration

Grzegorz Skoraczyński[1]

[1]University of Warsaw, Warsaw, Poland

The problem of automatic prediction of organic reaction yield and duration time is known to be challenging. In this work we study the ability of machine learning methods to solve this problem. Machine learning methods turned out to be very efficient and popular in various applications where the goal is to predict the answer from multidimensional vector of features. To this aim the huge amount of data is processed and the appropriate model is created.

For a problem of predicting yield and duration time the space of features is chosen to be the widely used space of numerical descriptors of reaction substrates nad products. The obtained classification accuracy were ca. 65% for reaction yield and ca. 75% for reaction duration time. Moreover this rather pessimistic outcome cannot be significantly improved. We provide the proof of this statement by approximating the upper bounds for accuracy of Bayes classifier, which can be interpreted as maximal accuracy achievable by ML algorithms. Obtained results were at levels of 70% and 80% for reaction yield and duration, respectively.

These results fully explain the non-triviality of the prediction task. We also demonstrated potential hardness of improving these result. We designed a dedicated tree kernel for SVM classifier. We used atom-mapping algorithm in order to provide as an classifier input more precise information of changing molecules topology. It turned out that accuracy of the classifier in this extended approach was at the same level.

(In)Stability of protein-coding genes’ 5’ end overlap

Wojciech Rosikiewicz[1,2], **Yutaka Suzuki**[3], **Sheng Li**[2], **Izabela Makałowska**[1]

[1]Adam Mickiewicz University, Poznan, Poland

[2]The Jackson Laboratory for Genomic Medicine, Farmington, USA

[3]The University of Tokyo, Chiba, Japan

For a long time, gene overlap was considered to be rather uncommon, but nowadays more and more of different types of overlapping genes, depending on their position and the transcription direction, is reported in diverse species. Gene overlap is often considered as a fixed feature, that can be assigned to a specified gene pair. Here, using the information of the alternative transcription start sites (TSS) in 73 human and 10 mouse organs, tissues and cell lines, we have investigated the (in)stability of the 5’ ends gene overlap among hundreds of identified gene pairs. We show, that the single gene pair, identified as overlapping in one library, may use a different set of alternative TSS in different libraries. In the case of about 50% of identified genes, it results in the tissue-specific transcription initiation of genes without overlap. It suggests that gene overlap may either play an important regulatory function, for example via mechanisms like transcriptional interference, or that it is just a side effect of the alternative transcription start sites usage. We investigate which of these possibilities is more likely by the integrative analysis of transcriptomic data, like TSS-Seq and RNA-Seq, combined with epigenomic data analysis, including seven types of histone modifications, strengthen by the ChIP-Seq based analysis of RNA Polymerase II activity. Finally, we investigate the allele-specific expression of overlapping genes in the context of DNA methylation “epialleles”, determined using a new method based on ultra-long sequencing reads from Oxford Nanopore Technologies. The key results of the project were stored in a database, accessible under http://overgenedb.amu.edu.pl.

Session B3-2 Experienced researchers’ talks

Photoinduced Structural Rearrangements in Proteins Involved in Insulin Release. Computer and Bioinformatics Modeling

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

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

Type 2 diabetes (T2D) affects millions of people worldwide. To control blood glucose level sulfonylurea drugs are prescribed. Unfortunately, the peak concentrations of the drugs are often at variance with demand for pancreatic beta cells activation governed by eating habits of TD2 individuals. Recent progress in light activated drugs [1] offers new ways of more precise time control of insulin release. In order to understand better structural determinants of early steps in this process, we modeled two protein systems: Kir6.1/SUR1 and EPAC2. Newly developed photoactive sulfonylurea derivatives [2] were docked to both proteins. Binding pockets were analyzed in all details. In particular, all residues affected by the trans-cis photoisomersation of the azobenzene moiety were scrutinized using the pocket volume analysis and themolecular dynamics simulations in ground and excited states. The analysis was focused on point mutations linked with neonatal diabetes. Our work demonstrates that computer modeling may help to understand structural aspects of critical physiological processes.

[1] R. Brazil, The Pharmaceutical Journal 9 FEB 2017.

[2] Broichhagen J, Schönberger M, Cork SC et al. Optical control of insulin release using a photoswitchable sulfonylurea. Nat Commun 2014; 5:5116. doi: 10.1038/ncomms6116

This work is supported by National Science Centre, Poland grant no. 2016/23/B/ST4/01770 and ICNT

Bioinformatics discovery of novel kinase families — unexpected biology and biochemistry in a well-known enzyme superfamily

Krzysztof Pawłowski[1]

[1]Warsaw University of Life Sciences SGGW, Warsaw, Poland

Protein kinases are an extremely important group of enzymes that mediate signalling by phosphorylating various substrates in the cell. Despite longtime intensive research, novel kinases and kinase families still happen to be found. We employ bioinformatics tools for remote homology detection and structure prediction to identify suchnovel families.Comparing a number of novel kinase families found by us and by others in recent years (e.g. FAM20, COTH,SELO, FAM69), and including some yet unpublished families, we estimate difficulties encountered in identifying such distant members of diversified protein superfamilies. We discuss some novel very distant kinase and pseudokinase families, with additional difficulty lying in partly lacking active sites. We also indentify a case of circularly permuted elements of the classic kinase domain and non-kinase insertions within the kinase domain. The most exciting and unexpected novel kinase-like family that we have found to-date is the ubiquitous eukaryotic-bacterial SELO/YdiU family that we highlighted as unusually well-conserved family of putative enzymes. Our recent experimental results, from a joint collaborative project (unpublished, in revision) show that SELO, although having a protein kinase-like three-dimensional structure, utilizes ATP in a manner different from all other kinases, for AMPylation (adenylylation) of protein substrates.We also present data showing that this function is conserved in SELO in evolution (bacteria, yeast, humans) and important for response to oxidative stress. We argue that search for distant members of established enzyme families can bring discoveries such as variations of known functions or novel, unexpected functions.

SURPASS: Low-Resolution Coarse-Grained Protein Model

Dominik Gront[1], **Aleksandra Elżbieta Dawid**[1], **Andrzej Koliński**[1]

[1]University of Warsaw, Warsaw, Poland

The recently published SURPASS (Single United Residue per Pre-Averaged Secondary Structure fragment) model has been proposed to fatilitate fast modeling of large proteins. The design of the model is unique and strongly supported by the statistical analysis of structural regularities characteristic for protein systems. Coarse-graining of protein chain structures assumes a single center of interactions per residue and accounts for preaveraged effects of four adjacent residue fragments. Knowledge-based statistical potentials encode complex interaction patterns of these fragments. Using the Replica Exchange Monte Carlo sampling scheme and a generic version of the SURPASS force field we performed test simulations of a representative set of single-domain globular proteins. The method samples a significant part of conformational space and reproduces protein structures, including native-like, with surprisingly good accuracy.

Poster: P24

Understanding unmapped reads using *Bos taurus* whole genome DNA sequence

Joanna Szyda[1,2], **Magda Mielczarek**[1,2]

[1]Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2]National Research Institute of Animal Production, Balice, Poland

Alignment to the reference genome is one of the most important aspects of studies based on whole genome DNA sequence. It is an intermediate step of all studies using Next Generation Sequence data. Further steps down typical bioinformatic pipelines focus on short reads, which were successfully (i.e. uniquely and with high confidence) aligned to the reference genome, while unmapped reads are disregarded. Nevertheless, such unmapped reads may carry valuable biological information, which can help in finding new genomic regions previously excluded from the reference genome, or to identify specific characteristics of individual-genomes. Therefore, in this study, we aimed to explain possible reasons of non-aligned reads. The study material consisted of whole genome DNA sequences of 48 Brown Swiss bulls, sequenced in paired-end read mode on the Illumina HiSeq 2000, with a genome averaged coverages varying between 8-28. Short reads were aligned to the reference genome. Then, we selected unmapped reads of good overall quality expressed by the read average quality threshold of 20. Such edited reads were compared between bulls in order to identify sequences of high similarity between all individuals. Furthermore, BLAST was used to find location of those reads, based on the NCBI nucleotide data base.

The research was supported by the European Union's Seventh Framework Program for research, technological development and demonstration under grant agreement 289592 - Gene2Farm and by Poznan Supercomputing and Networking Centre.

Dnaasm — de-novo assembler for second and third generation sequencing data

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

[1]Warsaw University of Technology, Warsaw, Poland

We developed the de-novo DNA assembler called 'dnaasm'. Our algorithm uses the relative frequency of reads to properly reconstruct repetitive sequences (tandem repeats), which are longer than the insert size of paired-end tags. Our results are about 5% better in terms of number of contigs and N50 statistic over other known solutions. Tandem repeats could also be restored, if only single-read sequencing data is available. The algorithm is based on Pevzner graph and is able to analyse the output from second generation sequencers, like Illumina. Moreover we implemented the scaffolding module, able to use long DNA reads with high level of errors, produced by third generation sequencers, like MinION Oxford Nanopore. The long DNA reads are used to determine the order of contigs produced by our assembler. We developed the application using 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 and used in both: simulated and real data. We use this application to create tapeworm (\emph{Hymenolepis diminuta}) draft genome. Our application allows to hybrid DNA assembly, where Illumina paired-end tags of length 250bp, Illumina mate-pairs with insert 2kbp and MinION with average length 7kbp were used together. Final value of N50 was equal to 2537 kbp, the number of sequences greater than 1000 bp is 719 and the sum of the sequences increased to 177.348 Mbp. For our knowledge it is the best genome draft for this species.

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 under MIT license.

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], Janusz Bujnicki[1]

[1]International Inst. of Molecular and Cell Biology, Warsaw, Poland

Macromolecular complexes composed of proteins and nucleic acids play fundamental roles in many biological processes. Structures of some of these complexes have been experimentally determined, 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 disordered 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.

Poster: P03

POSTERS

P02: Predicting likelihood of organic molecules by structure based Markov random field

Grzegorz Skoraczyński[1], Błażej Miasojedow[1], Sara Szymkuć[2], Ewa Gajewska[2], Bartosz Grzybowski[2], Anna Gambin[1]

[1]University of Warsaw, Warsaw, Poland

[2]Polish Academy of Sciences, Warsaw, Poland

Computer aided retrosynthesis is a forward-looking method of automating the process of chemistry development. The challenging open problem is to restrict a space of potential reactants to those which are feasible. There are many solutions proposed and we focus on assessing the probability of a given organic molecule based on its structure. Our goal is to create an oracle tool which can answer whether a given molecule is possibly existing, e.g. is stable and there exist molecules with similar structure. To this aim we employed a set of over 200 motifs schemes, which were carefully curated by experts. Pattern is usually a generic schema of functional group or layout of groups. Moreover, we created a set of descriptors characterizing a molecule by a specific motifs layout. To grasp a structure of interactions between descriptors we applied a Markov Random Field model. The model infers the probability distribution over graphs of interactions between descriptors. The structure of graph and the strength of interactions between vertices are learned by sparse Machine Learning techniques. We estimated parameters of the model from over 6 million entries database. Thanks to this model we can easily and fast evaluate a score of molecule with given structure. This can help to restrict the size of space explored during computational retrosynthesis.

P05: The impact of selecting reference sample set on CNV callers results

Wiktor Kuśmirek[1], **Robert Nowak**[1], **Tomasz Gambin**[1]

[1]Warsaw University of Technology, Warsaw, Poland

There are over 25 tools dedicated to detect CNV based on read depth. Reported applications are composed of the several steps, which are processing input data in the similar idea. Firstly, the depth of coverage is counted for each exon in each sample. Then, normalization process is applied, which estimates depth of coverage in terms where CNVs do not occur. Finally, raw and normalized depth of coverage are compared and segmentation process is applied, which produce resultant set of CNVs. However, some CNV callers use specified algorithms to designated reference sample set before read depth normalization stage while others CNV callers use all samples as a reference set. Herein, we used whole-exome sequencing data from the public domain to evaluate another types of reference sample set selection algorithms. Initially, number of mapped reads in each exon for each sample were counted. Subsequently, we discarded outstanding exons based on read depth, length of the target, GC content etc. Then, for each sample we designated a set of reference samples using different methods and parameters. Finally, we used chosen samples in normalization and CNV calling processes. The described process was repeated several times for various methods of reference selection and various CNV callers with another parameters. The presented methods of selecting reference sample set before normalization process could improve the results of detection CNVs in the human genome. In particular, the number of false positives calls could be significantly decreased by proper selection of reference sample set.

P06: To uncover what is hidden — studying amyloidogenic propensity

Natalia Niedzielska[1], Marlena Gąsior-Głogowska[1], Michał Burdukiewicz[2], Małgorzata Kotulska[1]

[1]Wroclaw University of Science and Technology, Wroclaw, Poland

[2]University of Wroclaw, Wroclaw, Poland

Amyloidogenity of peptides is directly bound with many neurodegenerative diseases. Therefore, amyloidogenic regions are responsible for amyloid formation and aggregation. In our research we attempt to identify the potential amyloidogenic amino acids sequences by Attenuated Total Reflection–Fourier Transform Infra-Red (ATR-FTIR) spectroscopy. The study then aims to validate results from ATR-FTIR technique by different bioinformatics methods. The experiment was performed on the twenty-four synthetic hexapeptides (CASLO company). We conducted infrared spectroscopy experiment by using air-dried peptide films. We performed on the peptides bioinformatical computational methods such as: TANGO, FoldAmyloid, AGGRESCAN, Waltz, PASTA 2.0, FISH Amyloid, APPNN and Amylogram. In results it was observed that the main structural features of amyloid aggregates are anti-parallel β-sheets. Amyloid fibrils have a spectral signature clustering between 1611 and 1630 cm−1. The presence of the characteristic relatively intense absorption band in the infrared spectra easily distinguishes the aggregated proteins. We found that some peptides were incorrectly classified as non-amyloid in literature. In results obtained by bioinformatical methods we can observed big threshold of positive typing (15, 13,12 for, subsequently, APPNN, FISH, Amylogram and 2, 3, 4 for TANGO, PASTA 2.0,WALTZ). However, by summing up the number of assignments for given sequences, it is possible to determine the most frequently typed ones. It is highly probable that these sequences actually form as amyloids. The ATR-FTIR experimental method allows comparable results for the bioinformatic methods. Decisive advantages of bioinformatic methods are: lower cost, time consumption and greater repeatability.

P07: Graph comparison by decomposition to subnets

Bartłomiej Szawulak[1], **Piotr Formanowicz**[1]

[1]Poznan University of Technology, Poznan, Poland

Graph comparison analysis is an important problem, especially in the context of analysis of graph-based models of biological phenomena. Unfortunately, most of algorithms for graph isomorphism problem return only a single value of the similarity. This might cause issues when graph-based models of biological systems are compared.

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

P10: Estimating probabilistic context-free grammars for proteins using contact map constraints

Witold Dyrka[1], **François Coste**[2], **Hugo Talibart**[2]

[1]Wroclaw University of Science and Technology, Wroclaw, Poland

[2]Université de Rennes, Inria, Rennes, France

Learning language of protein sequences, which captures non-local interactions between amino acids close in the spatial structure, is a long-standing bioinformatics challenge, which requires at least context-free grammars. However, complex character of protein interactions impedes unsupervised learning of context-free grammars. Using structural information to constrain the syntactic trees proved effective in learning probabilistic natural and RNA languages. In this work, we establish a framework for learning probabilistic context-free grammars for protein sequences from syntactic trees partially constrained using amino acid contacts obtained from wet experiments or computational predictions, whose reliability has substantially increased recently. Within the framework, we implement the maximum-likelihood and contrastive estimators of parameters for simple yet practical grammars. Tested on samples of protein motifs, grammars developed within the framework showed improved precision in recognition and higher fidelity to protein structures. The framework is applicable to other biomolecular languages and beyond wherever knowledge of non-local dependencies is available.

This research has been partially funded by National Science Centre, Poland [grant no 2015/17/D/ST6/04054] and was supported by the E-SCIENCE.PL Infrastructure. Computational experiments have been partially carried out using resources provided by Wroclaw Centre for Networking and Supercomputing (http://wcss.pl) [grant no 98].

P12: The role of YRNA-derived small RNAs in atherosclerosis development and progression — modeled and analyzed using stochastic Petri nets

Agnieszka Rybarczyk[1]

[1]Poznan University of Technology, Poznan, Poland

Non-coding RNAs 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.

P13: *De novo* genome assembly for third generation sequencing data

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

[1]Warsaw University of Technology, Warsaw, Poland

The second generation sequencing techniques opened doors to further research on a world scale, because the cost of DNA sequencing dropped significantly. However, the second generation sequencing technology has some drawbacks, mainly short read length. In 2017 the new devices, that use real-time sequencing started to be available. This approach, called "the third-generation sequencing" achieve read length of 20kbp and error rate about 15%. As a consequence of this process new DNA assemblers were developed. In this article we propose an implementation of Overlap Graph-based de novo assembly algorithm for third-generation sequencing data. The proposed method involves graph algorithms and dynamic programming, optimized using a MinHash filter. The solution has been tested on both simulated and real data of bacteria obtained from Oxford Nanopore MinION sequencer. The algorithm is included in "OLC" module of the dnaasm de novo assembler. Dnaasm application provides command line interface as well as web browser-based client. 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.

P14: Implementation of a hybrid algorithm for generating scaffolds

Wiktor Franus[1], Wiktor Kuśmirek[1], Robert Nowak[1]

[1]Warsaw University of Technology, Warsaw, Poland

The second generation sequencing methods produce high-quality short reads, which are assembled into contigs by DNA assemblers. Due to the fact that length of a single read is limited to 500bp, the contigs are far shorter than chromosomes. Generating longer contigs is a main effort of computer scientists in this area. The most popular approach for contig extension is to use pair-end tags, however this method does not always produce best possible results because distances between two paired reads are limited to 8kbp. An another, recently published approach appears to be free of this disadvantage. Contig linking is performed using reads of lengths reaching 20kbp. The above-mentioned reads can be obtained as a result of third-generation sequencing. An existing implementation of this algorithm appears to be time and memory demanding for larger genomes. We propose a new optimization of this algorithm based on Bloom filter and extremely memory-efficient associative array. Our improved implementation remarkably exceeds the previous one in terms of time and memory consumption. The updated algorithm has been tested on real data of bacteria, yeast and plant genomes. The optimized algorithm, provided as a library, is a part of the dnaasm de novo assembler. The library has been created using C++ programming language, Boost and Google Sparse Hash libraries. Both web browser-based graphical user interface and command line interface are provided. 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.

P15: Genome-wide analysis of single nucleotide polymorphism underlying feet and leg disorders for Austrian Braunvieh cattle

Barbara Kosińska-Selbi[1], Joanna Szyda[1,2], Magdalena Frąszczak[1], Tomasz Suchocki[1,2], Christa Egger-Daner[3], Hermann Schwarzenbacher[3]

[1]Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2]National Research Institute of Animal Production, Balice, Poland

[3]ZuchtData EDV-Dienstleistungen GmbH, Vienna, Austria

The aim of this project is to analyse single nucleotide polymorphisms (SNPs) to identify mutations underlying feet and leg disorders of Austrian Braunvieh cattle. 985 cows were genotyped using the GeneSeek® Genomic ProfilerTM HD oligonucleotide microarray, resulting in 76 932 SNPs available for each individual. Those polymorphisms are located relatively evenly on all chromosomes (from 4 514 SNP on chromosome 1, to 77 SNPs on chromosome Y). The phenotypic data contains information on the total number of hoof disorders scored until 100th and 300th day of lactation (discrete values from 0). The PLINK software (v1.07) was used for data editing. The GCTA software (v.1.94) was used to estimate SNP effects by fitting the following mixed linear model (MLM): $y = a + bX + g + e$, where y is the breeding value, a is the mean term, b is the additive fixed effect of the candidate SNP to be tested for association, X is the SNP genotype design matrix coded as 0, 1 or 2, g is the random polygenic effect and e represents a random residual. The normal distribution is preimposed on the polygenic effect with $g \sim N(0,G)$, where G corresponds to the covariance matrix between SNPs multiplied by the polygenic variance. Residuals are iid, expressed by $e \sim N(0,R)$, with R being an identity matrix multiplied by the residual variance. The mean value for the breeding value (estimate for data on the total number of hoof disorders scored until 100th day of lactation) is equal to 0.4917 and standard deviation equal to 0.0049. The mean value for the breeding value (estimate for data on the total number of hoof disorders scored until 300th day of lactation) is 0.3993 and standard deviation equal to 0.0376.

P18: Unraveling the genetic background of clinical mastitis in dairy cattle using WGS

Joanna Szyda[1,2], Magda Mielczarek[1,2], Magdalena Frąszczak[1], Giulietta Minozzi[3], John Williams[4], Katarzyna Wojdak-Maksymiec[5]

[1]Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2]National Research Institute of Animal Production, Balice, Poland

[3]University of Milan, Milan, Italy

[4]University of Adelaide, Adelaide Australia

[5]West Pomeranian University of Technology, Szczecin, Poland

Mastitis is an inflammatory disease of the mammary gland, which has recently become one of the most important diseases of the dairy sector, mainly due to high economic importance and increased awareness of animal welfare. The aim of this work was to characterize links between single base pair (SNPs) and structural (CNVs) variations and the incidence of clinical mastitis. Using information from 16 Holstein-Friesian cow half-sib pairs (32 animals in total) we searched for genome regions differing between a healthy (no incidence of clinical mastitis) half-sib and a mastitis-prone (multiple incidences of clinical mastitis) half-sib. The most interesting outcome emerged when CNV regions deleted in a sick cow, but present in its healthy half-sib were considered. 191 such regions differing in at least nine sib-pairs were observed. Those regions overlapped with exonic sequences of 46 genes. Among them, based on literature evidence, APP (BTA1), FOXL2 (BTA1), SSFA2 (BTA2), OTUD3 (BTA2), ADORA2A (BTA17), TXNRD2 (BTA17), NDUFS6 (BTA20) exhibited potential causal influence on clinical mastitis. There were also SNPs with differential genotypes between a healthy and a sick sib. Among 17 SNPs overlapping among at least eight half-sib families, three were located introns of genes: MET (BTA04), RNF122 (BTA27) and WRN (BTA27). Concluding, in this study we observed that structural polymorphisms play an important role in susceptibility to clinical mastitis and sequence deletions exhibit more severe consequences on reducing resistance against mastitis, than sequence duplication on increasing resistance against the disease.

Acknowledgements: The research was supported by the European Union Seventh Framework Programme through the NADIR (FP7-228394) project, by the Polish National Science Centre (NCN) grant 2014/13/B/NZ9/02016. Computations were carried out at the Poznan Supercomputing and Networking Centre.

P19: Analysis of cancer genomic sequencing data by correlations of gene expression footprints

Mateusz Kania[1], **Paweł Sujak**[1], **Andrzej Polański**[1]

[1]Silesian University of Technology, Gliwice, Poland

Our research regards the study on genetic background of two cancers, Thyroid carcinoma and Glioblastoma multiforme. It is mainly focused on identifying a network of genes important in disease development. We modified the approach of methodology published by Pongor et al. In his paper the expression signature is constructed for linking the genotype of cancer patients to survival. In our approach we cluster expression signatures to obtain information on the networks of genes important for cancer development. We have used files from the TCGA DB, containing data of mutated genes and their expression. We choose 1/5 genes with the highest mutation frequency. For each of these, patients were divided into groups with and without the mutation. Utilizing gene expressions, multiple Kolmogorov-Smirnov tests were performed. Bonferroni correction was applied to reduce false discovery rate. As a result, we received statistically significant genes. We called it expression footprint. A footprint of genes was tested against each other to identify similarities between them. We created a matrix of related genes. Those were divided into clusters which were described by features of selected clusters, based on AmiGO DB. Groups of mutated genes were linked together by correlated expression footprint which allowed us to study the characteristic features of gene clusters. Using our methodology, we identified unique properties of clusters, described by gene ontology terms. Among few, there were: neuron development and differentiation, neuroogenesis or regulation of signal transduction.

Pongor, L. et al. Genome medicine 7.1 (2015): 104

P20: The influence of selection at the amino acid level on the usage of synonymous codons

Paweł Błażej[1], Dorota Mackiewicz[1], Małgorzata Wnętrzak[1], Paweł Mackiewicz[1]

[1]University of Wroclaw, Wroclaw, Poland

The effectiveness of protein translation is usually considered as the main selectional factor influencing the codon usage. However, the biased usage can also be a by-product of a general selection at the amino acid level interacting with nucleotide replacements. To evaluate the validity and strength of such an effect, we superimposed 3.5 billion unrestricted mutational processes on the selection of nonsynonymous substitutions based on the differences in physicochemical properties of the coded amino acids. Using a modified evolutionary optimization algorithm, we determined the conditions in which the effect on the relative codon usage is maximized. We found that the effect is enhanced by mutational processes generating more adenine and thymine than guanine and cytosine, as well as more purines than pyrimidines. Interestingly, this effect is observed only under an unrestricted model of nucleotide substitution, and disappears when the mutational process is time-reversible. Comparison of the simulation results with data for real protein coding sequences indicates that the impact of selection at the amino acid level on synonymous codon usage cannot be neglected. Furthermore, it can considerably interfere, especially in AT-rich genomes, with other selections on codon usage, e.g., translational efficiency. It may also lead to difficulties in there cognition of other effects influencing codon bias, and an overestimation of protein coding sequences whose codon usage is subjected to adaptational selection.

P21: Predicting clinical endpoints of breast cancer patients by integrating clinical and molecular data

Aneta Polewko-Klim[1], Witold Rudnicki[1]

[1]University in Bialystok, Bialystok, Poland

Motivation. The goal of the study was improving prediction of the clinical endpoint of breast cancer patients (CEBCP), by integrating data from different sources. We used clinical descriptors (CD), gene expression (GE) and somatic copy number aberrations (CNA) collected in the METABRIC dataset.

Method. We built predictive models for the clinical endpoints with the help of Random Forest classification algorithm using variables that are identified as relevant by five different feature selection methods. For robustness of results entire modelling procedure was performed in the cross-validation scheme that was repeated multiple times. Various methods for combining information from many sources were tested. The best method involves generating separate models for CD, GE and CNA datasets, and then extending the CD model with results of CD and GE models.

Results. The best results among models built using single source of information were obtained for CD dataset, MCC=0.32. The models built on GE and CNA datasets achieved MCC=0.25 and MCC=0.19, respectively. Integration of information from molecular datasets improves the predictive power of models. Adding the synthetic variable representing prediction of RF model built on CNA dataset increases MCC by 0.03, adding variable built using GE dataset improves MCC by 0.05 and adding both improves MCC by 0.06. Interestingly, sets of variables identified as most relevant by different feature selection methods have little common elements, nevertheless. Nevertheless, RF models built using these diverging sets of variables achieve nearly identical predictive power. What is more, while these method-specific sets of variables don’t give much insight into biological background, their combination reveals clear information of molecular pathways.

P22: Predicting protein architecture with Bayesian Networks

Konrad Bednarek[1], **Dominik Gront**[1]

[1]University of Warsaw, Warsaw, Poland

Protein secondary structure is a very important step in predicting protein three-dimensional structure. For instance, the recently proposed coarse grained SURPASS model uses secondary structure information as its only input to sample topologies accessible to a polypeptide chain. Here we propose a Bayesian network approach to predict protein secondary structure in an extended alphabet where additional symbol differentiates between edge and inner strands. Replica Exchange Monte Carlo method is used to sample posterior distribution. Prior probabilities have been derived from known protein structures. The method can also estimate probabilities for strand pairing. The method will facilitate accurate predictions by the SURPASS approach.

P23: *Legionella* effector kinases

Marcin Gradowski[1], Krzysztof Pawłowski[1]

[1]Warsaw University of Life Sciences, Warsaw, Poland

Legionella, the notorious intracellular pathogen, settles down inside the cell of its eukaryotic host and victim, and by using an imposing set of 300+ effector proteins establishes itself a comfortable niche. These effectors, delivered into the host cell, massively rewire the signalling pathways. Although most of the effectors are still little understood, several surprising distant homologues of eukaryotic signalling proteins were identified among them in recent years. Here, we focus on the proteinkinase-like (PKL) proteins among Legionella effectors. We present a bioinformatics analysis of the known PKL effectors, e.g. MavQ, LegK1-4, LepB, as well as a novel PKL effector family, and compare them to non-effector Legionella kinases and to human host kinases. Using sequence and structure data, we discuss the evolutionary relationships of the novel enzyme families to the known distant relatives. We show how substrate specificity may have evolved to allow bacterial proteins (own or acquired) specifically hit intracellular eukaryotic targets. Using the novel *Legionella* PKL effector family as an example we show how bioinformatics algorithms for three-dimensional structure prediction and remote homology detection help in finding novel enzyme families. This appears to be specially the case for intracellular pathogens where evolutionary arms race between the pathogen and the host is intense, and opportunities for inter-kingdom horizontal gene transfer are open.

P25: Bioshell suite v.3.0: tests and applications

Joanna Magdalena Macnar[1], Natalia Anna Szulc[1], Aleksandra Elżbieta Dawid[1], Dominik Gront[1]

[1]University of Warsaw, Warsaw, Poland

BioShell is a software toolkit for structural bioinformatics that has been developed for more that ten years. Its newest version, implemented in C++0x brings several new applications and an exhaustive set of tests and benchmarks. The new applications are devised to aid molecular modelling tasks such as docking, protein structure prediction and analysis. The new version of the package has been released on Apache 2.0 license

P26: Induced fit docking with AutoDock and Modeller

Alicja Płuciennik[1], **Marcin Pacholczyk**[1]

[1]Silesian University of Technology, Gliwice, Poland

Structure based virtual screening is commonly used in computer aided drug design pipelines. Using molecular docking one can scan large databases of small molecules against potential molecular targets (proteins). Effective protein-ligand docking algorithm should present fair trade off between speed and accuracy to be useful in high throughput virtual screening. We present novel approach to modelling of induced fit effect. Our Induced Fit Docking (IFD) protocol models mutual adaptations of protein receptor to small molecule ligand upon binding. While offering better accuracy than docking to rigid target, IFD requires less computational effort than more complex techniques like molecular dynamics (MD).The protocol starts with docking with softened potential (so called soft docking), and proceeds through selective optimization of side chains and docking into ensemble of optimized protein receptors. IFD was implemented using popular tools freely available for academic research (AutoDock 4 and Modeller).

P28: Homology modeling and n-gram analysis in amyloidogenecity prediction

Jakub Wojciechowski[1], **Małgorzata Kotulska**[1]

[1]Wroclaw University of Science and Technology, Wroclaw, Poland

Many neurodenerative diseases including Alzheimer’s disease are closely connected with occurrence of plaques formed by protein aggregates called amyloids. Although they become object of great interest of researchers around the world it is not entirely clear what causes misfolding and aggregation of this proteins. Plenty of studies shown that regions in proteins called hot spots are crucial for this process to occur thus many experimental studies aimed to identify those regions were conducted. However currently available experimental techniques are expensive and time consuming thus it is not possible to use them for large amount of samples. To overcame this problem plenty of computational methods of amyloidogenicity prediction were proposed. Although some of them achieve very good results, it is often difficult to interpret them. In order to provide more insight into structure of hot spots, we proposed amyloidogenecity method based on homology modeling. We show that using structure of amyloid fibril from yeasts as a template we can distinguish between amyloidogenic andnon-amyloidogenic fragments. We also checked if other methods of protein structure prediction can be used to predict amyloidogenecity. Finally we analyzed occurrence of most informative n-grams analyzed by authors of AmyloGramin different types of amyloid fibrils.

P29: MotifLCR: A new method for clustering low complexity regions

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

[1]University of Warsaw, Warsaw, Poland

[2]Silesian University of Technology, Gliwice, Poland

[3]Polish Academy of Sciences, Warsaw, Poland

Low complexity regions (LCRs) are fragments of proteins sequences with low diversification of amino acids. Such fragments can be found in proteomes more often than by chance. About 10% of human proteins contain LCRs [1] and these may often play very important role for appropriate structure and function of proteins [2]. One can find a plethora of methods to find LCRs, however there is no programme to compare LCRs. We want to solve this lack of tools by propose a new method designed specifically for searching for similar low complexity regions in protein sequences. The first step of algorithm looks for repeats in sequences. It uses a sliding window which is changing iteratively scanning the sequence multiple times. The method assigns sequences to clusters represented by corresponding type of repeats. After that the algorithm builds PSSMs. Each of such PSSM represents specific cluster. This step 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 junks in input sequences are removed before processing. In the second step sequences that are not assigned to any cluster are compared with the existing clusters and assigned to the most similar ones. MotifLCR also creates clusters of LCRs composed of two or more fused low complexity fragments.

References:

1. Núria Radó-Trilla and MMar Albà. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins.BMC Evolutionary Biology 2012

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. JBC 2015

P30: Graph construction algorithm for chemical compounds representation

Alexander Antkowiak[1], Piotr Formanowicz[1]

[1]Poznan University of Technology, Poznan, Poland

A basic problem of graph construction on the basis of predefined vertex degrees is well known in graph theory. Chemical compounds can be represented by graphs as a structural formula representation. If we consider a problem of constructing structural formulas of chemical compounds that are more complex, basedon information of the number of atoms and their valencies, it is moreproblematic to obtain a solution. We have vertices that are equivalent to atomsand degrees of vertices that are equivalent to valencies. The edges representthe chemical bounds but if we need a more complex model of a chemical compound,we should consider for examplemultigraphs, where parallel edges are possible. In such a case a more realistic model of a chemical compoundcould be possible and this could lead to a construction of more precise algorithms solving the problem of structural formulas reconstruction for a better representation of molecules in mass spectrometry.

P32: How to create forests by cutting trees

Michał Ciach[1], Anna Muszewska[2], Paweł Górecki[1]

[1]University of Warsaw, Warsaw, Poland

[2]Polish Academy of Sciences, Warsaw, Poland

Horizontal gene transfer (HGT), a process of acquisition and fixation of foreign genetic material, is an important biological phenomenon. Several approaches to HGT inference have been proposed. However, most of them either rely on approximate, non-phylogenetic methods or on the tree reconciliation, which is computationally intensive and sensitive to parameter values. We investigate the locus tree inference problem as a possible alternative that combines the advantages of both approaches. We present several algorithms to solve the problem in the parsimony framework. We introduce a novel tree mapping, which allows us to obtain a heuristic solution to the problems of locus tree inference and duplication classification. Our approach allows for faster comparisons of gene and species trees and improves known algorithms for duplication inference in the presence of polytomies in the species trees.

P33: Metagenome classification for environmental sample analysis

Jolanta Kawulok[1]

[1]Silesian University of Technology, Gliwice, Poland

Nowadays, not only the analysis of single genomes is developing rapidly, but also of the metagenomes-entire collections of genomes derived from a single location. Metagenome analysis has a large potential due to the fact that it is not necessary to isolate and culture organisms in the laboratory to study them. The goal of our current work is to develop new methods for analysis of metagenomic reads which come from the microorganisms living in a given place. In our research, we focus on supervised classification of metagenomic data, in particular on the taxonomic classification (where the goal is to determine the types of organisms) and the environmental classification. The latter case consists in determining the origin of an environmental sample, which is faster, while often sufficient, than to identify all the organisms living in the investigated place. In order to determine the origin of a sample, first we build separate databases of k-mers (i.e., all substrings in a sequence of the length k) from the metagenomic reads for each class¬¬ that represents a place with which we want to compare the investigated sample. Then, every read acquired from the given sample is compared with them. The similarity (the match score) between the query read and each class is obtained based on the number of nucleotides in the k-mer occurring both in the read and in the class. Finally, we analyse all the match scores and based on them, the studied sample is classified to the appropriate class. The results of our experimental validation proved the feasibility of our approach and its applicability for analysing metagenomic samples.

P34: Molecular selectivity of antiseptic detergents from gemini family

Mateusz Rzycki[1], Beata Hanus-Lorenz[1], Sebastian Kraszewski[1]

[1]Wroclaw University of Science and Technology, Wroclaw, Poland

The widespread use of antibiotics led to rapid evolution of bacteria resistant to several drugs. One of the most promising ideas is to provide a molecule that will be able to change the mechanical properties of bacteria membrane that will lead to its destabilization and will be possible to distinguish the bacteria from the rest of the environment. In current work we present the results coming from combined methodology using molecular dynamics and massive docking approaches for among others chlorhexidine and octenidine. Chosen molecules are gemini type detergents presenting the lowest effective concentrations against a broad spectrum of bacteria. Supposed molecular mechanism of action consists on (i)molecule landing on the surface, (ii) incorporation within the membrane, and(iii) membrane emulgation. Expecting the important selectivity between eukaryotic and bacterial membranes we prepared all atom models for molecular dynamics study. One component PC membrane and three components bacterial membrane (80%PE, 15% PG, 5% CL) were evaluated once over 200ns, then obtained trajectory was used several times for massive docking. Expecting that thermal movements are important for local membrane defects formation being at the origin of ligand anchorning, we evaluated over 80'000 docked configurations predicting energetics of antibacterial process. This took only 3 days against many weeks if using molecular dynamics approach. We compared energy of landing subprocess between eukaryotic and bacterial membranes finding statistically important selectivity.The data unrevealed us also the energy of membrane incorporation correlating with effective concentrations of tested molecules determined experimentally. We believe our approach will fasten the research of new antiseptic agents.

P35: Guiding Rosetta abinitio protocol with backbone NOE data

Justyna Kryś[1], Daria Wultańska[1], Dominik Gront[1]

[1]University of Warsaw, Warsaw, Poland

Here we propose a novel approach that combines Rosetta template-free protein structure prediction method (_abinitio_ Rosetta protocol) with fragmentary NOE experimental data measured for a protein backbone into a novel approach for rapid determination of protein structures. The method modifies _ss_pair_ Rosetta energy term to guide conformational sampling towards the topology of beta strands that has been observed experimentally. The new method has been tested on a set of 25 real experimental cases measured by Montelione group.

P36: Modeling cytosine methylation and demethylation pathways

Karolina Kurasz[1], **Joanna Rzeszowska**[1], **Krzysztof Fujarewicz**[1]

[1]Silesian University of Technology, Gliwice, Poland

Background. Regulation of gene expression is a complicated process, which consists on very different factors. Discovery of the TET family proteins shed new light on pathways of DNA demethylation and posed new questions on the role of these enzymes and catalyzed by them cytosine modifications in cellular processes. Material and methods. The levels of methylated cytosine moieties and products of DNA demethylation pathway in DNA of five cultured human cell lines have been assessed. From those cell lines, DNA was used in the measurement of nucleotides modifications. The role of particular modifications in living cells is not clear, however knowledge of the mechanism(s) responsible for emergence of different types of modifications may be important for understanding their role in differentiation processes or development of cancer cells. However even precise assessment of the levels of different modifications is difficult as it needs large quantities of cells and complicated mass spectrometry methods. To understand mechanisms which regulate the levels of different cytidine modifications one should know also which enzymes from TET family are most efficient in particular reactions in different cell types and this is also experimentally very difficult problem. Results. We propose a simple mathematical model concerning methylation and known cytidine modification pathways which is able to predict levels of particular cytidine modifications and the role of particular TET enzymes in their creation in cells with known TET transcript levels.

This work was funded by the Silesian University of Technology under grant BKM/508/RAU1/2017/12

P37: Don't fool yourself — a robust protocol for application of Machine Learning in Life Science

Wojciech Lesiński[1], **Aneta Polewko-Klim**[1], **Krzysztof Mnich**[1], **Witold R. Rudnicki**[1]

[1]University of Bialystok, Bialystok, Poland

Motivation. Machine learning (ML) allows to change huge and messy datasets, which are obtained from modern molecular biology experimental methods, into predictive methods. Every month hundreds of papers using ML with are published, however, they are often irreproducible, or in a best case don't use data in a best possible way. In particular, the estimates of models' performance are often too optimistic and, more often than not, no confidence intervals for performance measures are given.

Methods. We have developed a protocol that allows to fully utilise data and obtain robust estimate of model performance, including estimate of confidence interval. N repeats of a full modelling procedure, involving optimisation of parameters and feature selection is performed K-times within cross-validation loop. The results from NxK tests are then used to derive robust estimate of the average model performance.

Experiment. Current study demonstrates the application of this protocol on several high-dimensional molecular datasets and differences with simpler designs. Datasets from neuroblastoma data integration challenge within CAMDA 2017 and drug toxicity challenge within CAMDA 2018 were used. The quality of results is estimated using Matthews correlations coefficient (MCC).

Results. Significant differences between various protocols are demonstrated. Two main findings are the extent of bias introduced by flawed design of the study and dependence of results on the composition of the test set. MCC inflated up to 0.4 was obtained for models obtained with wrongly executed feature selection in comparison with the robust procedure. A negative correlation between MCC obtained on training set and test set, as well as a large variance of test set performance on test sets is also demonstrated.

P38: Comparison of tools for estimation of cancer sample contamination

Kinga Leszczorz[1], Andrzej Polański[1]

[1]Silesian University of Technology, Gliwice, Poland

The increase of knowledge about carcinogenesis and expansion of new technology has caused development of new tools to analysis cancer genomic data. Significant stage of analysis this genomic data is to estimate purity of tumor tissue. By taking a sample from tumor tissue we take also a microenvironment of cancer so immune and cells of blood vessels, fibroblasts and also a stromal cells and sometimes the non-cancer cells taking by accident. This factor and another factor such as tumor clonality and ploidy can coused cancer genomic analysis more complex and interpretation of results more complicated. In our work we focused only on estimating admixture rate of tumor tissue. We used next generation sequencing data from TCGA atlas, to compare most common tools to estimate contamination of the cancer sample which use many different models. Some of them perform for example analysis approximation of tumor mixture and other estimate contamination in another way. The task of estimation of percentage of normal cells in the tumor sample brings a lot of interpretation and analytical problems. Algorithm give us different results so estimated value of contamination is unreliable. Additionally, very often stage containing this problem is very time-consuming due to long preparation data process. Therefore, there is a real necessity to improve existing algorithms, develop new tools or create program consisting all of them. Our comparisons and analyzes of approaches for estimating contamination of tumor tissue by normal cells DNA are useful for interpreting results of algorithms of estimating clonal structures of cancer cell populations.

P40: In search of primary plastids' origin

Filip Pietluch[1], **Przemysław Gagat**[1], **Paweł Mackiewicz**[1]

[1]University of Wroclaw, Wroclaw, Poland

The Archaeplastida supergroup is one of several main lineages of Eukaryota that clusters together glaucophytes (*Glaucophyta*), red algae (*Rhodophyta*) and green plants (*Viridiplantae*). The most characteristic feature of these three lineages is a photosynthetic organelle called primary plastid. It has been acquired via primary endosymbiosis. During this event a cyanobacterium was engulfed by a heterotrophic unicellular eukaryote, kept as an endosymbiont and finally transformed into a true cell organelle. All currently available data unambiguously confirm that scenario. However, there are still many questions left without an unequivocal answer, e.g. number of primary endosymbiosis or branching order of Archaeplastida lineages. We are trying to provide answers for these problems by performing detailed phylogenetic analyses. We used a concatenated set of 16S and 23S ribosomal RNA sequences chosen from 287 species of cyanobacteria and plants, representing a large diversity of the lineages. Initially, we performed multiple sequence alignment using R-Coffee. Phylogenetic analyses included sophisticated approaches, e.g. covarion model assuming that the rates of evolution can change on different branches and/or site-heterogeneous mixture (CAT) model assuming site-specific rates and profiles. Calculated trees generally support with significant confidence the monophyly of the Archaeplastida and basal position of *Glaucophyta* and Rhodophyta to *Viridiplantae*. The basal position of primary plastids in the cyanobacterial tree was recovered with significant confidence only with covarion models. To answer the remaining questions concerning the origin of plastids and Archaeplastida classification, more sequence data is required, as well as development of new phylogenetic methods.

P41: Application of Gaussian mixture models to clonality analyzes in cancer genomics

Katarzyna Sieradzka[1], **Andrzej Polański**[1]

[1]Silesian University of Technology, Gliwice, Poland

This work is focused on the Gaussian mixture models analyzes applied to cancer genomics datasets obtained from the TCGA data portal. First, the implemented Gaussian mixture decomposition tools were verified on the basis of the analysis of artificial data sets. This step was based on the generation of Gaussian mixture models and the use of the Mclust function in R environment. The purpose was to develop methods for quantifying quality of parameter estimation and to search for the relationship between the level of component overlap and the error of assessment of the mixture components. The Gaussian mixtures were generated with known degrees of coverage. After classification, there were found matching clusters and basing on them, there were counted distances which were the main element to count the errors of assessments. The real genomics data sets came from patients with the thyroid cancer. The main goal was to examine the clonal structure of this type of cancer. In the analysis there was a need to call somatic mutations and copy number variations, using MuTect2 and VarScan2. While calling CNVs there was also a need to use few another tools to make some transformations of them, eg. copyCaller and the Circular Binary Segmentation Algorithm. The major point of the clonality analysis was the usage of two tools for clonal analysis – SciClone and the mclust package from R environment. What is interesting, we made a conclusion about how the SciClone analysis differs from the analysis made by the Mclust function - in order to make the clonal analysis using the mclust package, there is a need to make some additional implementations which will take into account the tumor evolution.

P42: Single cell precision for discovering patterns in clonal simulations

Krzysztof Szymiczek[1], **Andrzej Polański**[1]

[1]Silesian Technical University, Gliwice, Poland

To reveal patterns and phenomena in clonal evolution, simulations have to reflect cell differentiation in population. Existing Gillespie-based solutions operate statistically, so the population of cells is described by measures and individuals are not perceptible. We built new software (CCES - Clonal Cancer Evolution Simulator) to perform simulations of clonal evolution with a resolution down to a single cell, allowing the observation of known patterns in clonal evolution. The heart of our simulator is a modified Gillespie-based algorithm with Tau-Leaping. Initial set of cells is subject to multiple Monte Carlo steps in which decision is taken about the further lot of cells. Each Tau-Leap reflects one iteration of the simulation, and the outcome of previous iteration is the input to next one. We shorten the Tau-Leap step time to a minimum, so each cell is subject to one Monte Carlo step in one cycle. In order to track individual cells, we developed several innovatory techniques to save the resources running the simulation. We use partial genome compression by cross-referencing largest nested common parts and we use duty-cycle based data-to-task allocation to speed-up the parallel computations. Result of the simulations covers the known observed regularities and phenomena like: clonal interference, Muller's ratchet, selection, genetic diversification and the power of selection pressure. We obtained 2-d images of the clonal evolution along with several statistics and distributions describing the clonal structure of the population. Each cell data and population measures are exported, and can be subject to further analysis using 3-rd party tools.

P44: Two protein import pathways into the cyanobacteria-derived, photosynthetic organelles of *Paulinella chromatophora*

Katarzyna Sidorczuk[1], **Michał Burdukiewicz**[1], **Paweł Mackiewicz**[1], **Przemysław Gagat**[1]

[1]University of Wroclaw, Wroclaw, Poland

Paulinella chromatophora is anarmoured filose amoebae of the supergroup *Rhizaria* that harbours two photosynthetically active bodies of cyanobacterial origin (chromatophores), acquired independently of primary plastids of the supergroup *Archaeplastida*,i.e. photosynthetic organelles of glaucophytes, red algae and green plants. Similarly, to primary plastids, chromatophores have lost many essential genes, and transferred substantial number of genes to the host nuclear genome via endosymbiotic gene transfer, including those involved in photosynthesis. To investigate how nuclear-encoded proteins are imported into *Paulinella chromatophores*, we performed detailed bioinformatic analyses, including machine learning methods such as T-distributed Stochastic Neighbor Embedding, multi-method phylogenetic studies, extensive Blast homology searches as well as searches for potential targeting signals and biochemical properties in plastid-targeted proteins. According to our model, small proteins cross the outer chromatophore membrane in vesicles of the endomembrane system while the large ones use a yet unidentified translocon. Next, independently on their size, proteins are pulled into the chromatophore matrix through a simplified archaeplastidian-like Tic complex. Depending on their function, they either reside in stroma or can be subsequently imported into the thylakoid lumen or thylakoid membranes based on Sec, Tat and Srp pathways that are also present in primary plastids. Our results indicate that Paulinella similarly to archaeplastidians evolved two pathways for nuclear-encoded, chromatophore-targeted proteins and the same pathways for thylakoid trafficking, emphasising similarities between both types of photosynthetic organelles.