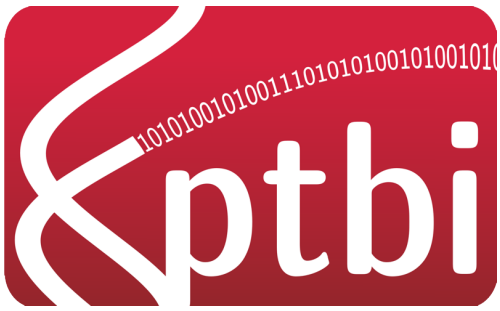
BIT19
Book of Abstracts

27–29 June 2019, Toruń, Poland



PROGRAM COMMITEE:

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

LOCAL ORGANIZING COMMITTEE:

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

Contents

Lectures

Rpart - clustering algorithm for big macromolecular data sets

Rafal Adamczak[1]

[1]Department of Informatics, Nicolaus Copernicus University

Clustering is a Machine Learning technique that involves the grouping of data points. It is widely used in the analysis and interpretation of molecular simulations for biological macromolecules, such as proteins and nucleic acids. Most of the existing clustering algorithms have time complexity of $O(n^2)$ what make them impossible to apply to big data sets. To make it feasible one may decrease number of data points by making microclusters, groups where only the most similar data points are joined. Two methods for microclusters will be presented one is based on profile hashing the other is making space partitioning by spheres (Rpart). Both methods have time complexity much lower the $O(n^2)$.

Email: `raad@is.umk.pl`

Mining molecular flexibility: novel tools, novel insights

Frédéric Cazals[1]

[1]Inria Sophia Antipolis - Méditerranée

This talk will review three recent developments fostering our understanding of molecular flexibility, with applications to structural and dynamical studies, as well as remote homology analysis.

The first development resides in the combined RMSD or RMSDComb. [1], a simple molecular distance mixing independent lRMSD measures computed with their own rigid motions. The combined RMSD is relevant to compare (quaternary) structures based on motifs defined from the sequence (domains, SSE), and to compare structures based on structural motifs yielded by local structural alignment methods. Illustrations of insights yielded by RMSDComb. will be given on (i) the assignment of quaternary structures for hemoglobin, and (ii) the analysis of conformational changes undergone by class II fusion proteins.

The second development [2] is a framework revealing the multiscale nature of motifs encoded within structural alignments returned by flexible aligners such as Kpax or FATCAT. The method combines in a bootstrap fashion ingredients from rigidity analysis (distance difference matrices), graph theory, computational geometry (space filling diagrams), and topology (topological persistence). The motifs identified inherently exhibit a hierarchical structure which sheds light on the trade-off between size and flexibility. They can also be combined to perform an overall comparison of the input structures in terms of combined RMSD. Illustrations will be provided on class II fusion proteins and hard structural alignments cases.

The third development leverages our structural motifs within a sequence-structure based method characterizing a set of functionally related proteins exhibiting low sequence identity and loose structural conservation [3]. We use our motifs to design hybrid sequence – structure based profile HMM characterizing protein families. These HMM are biased towards the structural properties encoded in motifs, and we show on class II fusion proteins that these HMM are able to retrieve homologous sequences beyond reach for sequence only based HMM.

All tools are available within the Structural Bioinformatics Library [4] (http://sbl.inria.fr), a highly modular C++/python library, which can be installed with conda and provides numerous jupyter notebooks to get started:

- Combined RMSD : https://sbl.inria.fr/doc/Molecular_distances_flexible-user-manual.html
- Structural motifs: https://sbl.inria.fr/doc/Structural_motifs-user-manual.html
- Hybrid HMM: https://sbl.inria.fr/doc/FunChaT-user-manual.html

[1] F. Cazals and R. Tetley. Characterizing molecular flexibility by combining lRMSD measures. Proteins, (in press), 2019.

[2] F. Cazals and R. Tetley. Multiscale analysis of structurally conserved motifs. Under review. 2019. BioRxiv preprint: https://www.biorxiv.org/content/10.1101/379768v1

[3] R. Tetley, J. Fedry, P. Guardado-Calvo, F. Rey, and F. Cazals. Hybrid sequence-structure based hmm models leverage the identification of homologous proteins: the example of class II fusion proteins. Under review. 2019. BioRxiv preprint: https://www.biorxiv.org/content/10.1101/379800v1

[4] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. Bioinformatics, 7(33):1–8, 2017.

Email: Frederic.Cazals@inria.fr

Machine learning and simulation methods for deciphering sequence-structure relationships in proteins

Jan Ludwiczak[1,2], Aleksander Winski[1], Birte Hernandez Alvarez[3], Vikram Alva[3], Maciej Malicki[4], Antonio Marinho da Silva Neto[1], Krzysztof Szczepaniak[1], and Stanislaw Dunin-Horkawicz[1]

[1]Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Poland
[2]Laboratory of Bioinformatics, Nencki Institute of Experimental Biology, Warsaw, Poland
[3]Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Tübingen, Germany
[4]Department of Mathematics and Mathematical Economics, Warsaw School of Economics, Poland

Proteins play a crucial role in nearly all biological processes ranging from cell division to its death. Despite the variety of the functions they perform, the chemistry of the protein world is rather simple, encoded by an alphabet of 20 amino acids. The order of amino acids in a protein sequence determines its structure, which, in turn, defines the function.

For a protein chain of a typical length, the number of possible amino-acid sequences is essentially infinite; however, only a tiny subset of this space was tested by nature and its majority remains unexplored. Computational protein design methods allow to "visit" the protein universe "dark matter" and thus create new proteins of previously unseen structures and functions.

In one of our projects, we attempt to design a "*pi*-helical coiled coil", a novel protein fold composed of two naturally occurring structural motifs – rare and unstable *pi*-helices and ubiquitous and stable coiled coils. Here, we present new bioinformatics tools that we have developed in the course of this project, namely neural network-based methods for accurate prediction of *pi*-helices and coiled coils in protein sequences and new protein design protocols, involving molecular dynamics simulations, enabling the exhaustive exploration of the sequence space.

Email: s.dunin-horkawicz@cent.uw.edu.pl

TBA

Chris T.A. Evelo[1]

[1]Department of Bioinformatics - BIGCaT NUTRIM School of Nutrition and Translational Research in Metabolism Faculty of Health, Medicine and Life Sciences

TBA

Email: chris.evelo@maastrichtuniversity.nl

SAXS data-driven modeling of RNA 3D structures

Pritha Ghosh[1], Michał J. Boniecki[1], Chandran Nithin[1], Gay Pauline Padilla-Meier[2], Grzegorz Chojnowski[3], Sean A. McKenna[2], Trushar R. Patel[4], and Janusz M. Bujnicki[1,5]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, ul. Ks. Trojdena 4, PL-02-109 Warsaw, Poland
[2]Department of Chemistry, University of Manitoba, Winnipeg, MB, Canada
[3]European Molecular Biology Laboratory (EMBL) Hamburg Outstation, c/o DESY, Notkestrasse 85, Hamburg 22607, Germany
[4]Alberta RNA Research and Training Institute, Department of Chemistry & Biochemistry, University of Lethbridge, 4401 University Drive, Lethbridge, Alberta T1K 3M4, Canada
[5]corresponding author

Majority of known RNAs exert their cellular functions in complexes with other molecules. Due to difficulties associated with experimental determination of high-resolution RNA structures, experimental data-aided computational modeling has become an important approach in generating high-quality theoretical models [Ponce-Salvatierra et al.2019]. The inherent flexibility of RNA molecules allows it to sample a large conformational space. This hints at the fact that its 3D structure is best represented by an ensemble of atomistic structures rather than a single structural model. The small angle X-ray scattering (SAXS) technique describes the distribution of electron density in a molecule and hence can be used to interpret the low-resolution envelope of a biomolecule. We have developed a computational workflow for SAXS data-driven modeling of RNA 3D structures and their ensembles. The workflow involves generation large sets of plausible conformations of target RNA with SimRNA (using a coarse-grained representation and a statistical potential to generate physically realistic structures) [Boniecki et al.2016]. These decoys are scored against experimental SAXS data using CRYSOL [Svergun et al.1995, Konarev et al.2006], followed by clustering, ensemble optimization [Tria et al.2015], and refinement with QRNAS. The workflow also allows use of data from other sources, such as information about RNA secondary structure from computational predictions or from experimental probing [Patel et al.2017]. Our method can also model 3D structures of RNA-protein and RNA-ligand complexes. In such cases, this method is used in conjunction with experimental restraints from other techniques. I will illustrate applications of our approach with case studies involving RNA molecules of different size.

Email: pghosh@genesilico.pl

Energy-dense diets leading to addiction affect brain gene expression and faecal microbiome content as well as microbiome gene activity.

Ilona Grabowicz[1], Julia Herman-Iżycka[2], and Bartek Wilczyński[2]

[1]Institute of Computer Science, PAS
[2]Institute of Informatics, Faculty of Mathematics, University of Warsaw

In our study we used highly-palatable and energy dense diets which are often consumed by humans leading to obesity, namely cafeteria and high-fat diet. We studied how cafeteria diet affected gene expression in three different areas of brains of mice. We observed that the genes of which expression levels correlated the most with animals' behavioral measurements, had merely weak expression changes between the control and cafeteria diet groups. These genes however seem to possess functions relevant for food seeking behavior and development of addiction. Although these behavior-correlating genes, as well as differentially expressed genes between diet groups, tend to be characteristic for each brain area, they cluster in the same topologically associating domains (TADs) irrespectively of a brain area (De Toma et al. 2018).

In the second part of the study we analysed the influence of a high-fat diet on the faecal microbiome and its gene expression enabled by the Next Generation Sequencing (NGS) technologies. With use of different NGS methods we concluded that long-term (14 or 28 days) use of high-fat diet lead to increase of abundance of two bacterial species: Mesoplasma florum and Lactobacillus salivarius. Moreover, we created a unique pipeline allowing for very high (83%) mappability of the NGS reads to the microbial genes. It allowed us to assess the changes in the microbial gene expression induced by high-fat diet. Functions of genes whose activity changed upon that diet included, among others, osmoprotection or lysine degradation. Moreover, we noticed that microbial gene expression profiles were highly individually specific meaning that each mice harbors its own microbiome gene expression fingerprint.

Email: ilona.grabowicz@gmail.com

Massive-scale structure and function predictions of human gut microbiome proteins for metagenomic applications

Tomasz Kosciolek[1,2], Douglas Renfrew[3], Vladimir Gligorijevic[3], Tommi Vatanen[4], Julia Koehler Leman[3], Ramnik Xavier[4], Rob Knight[2,5,6], and Richard Bonneau[3]

[1]Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[2]Department of Pediatrics, University of California San Diego, USA
[3]Flatiron Institute, New York, USA
[4]Broad Institute, Cambridge, MA, USA
[5]Department of Computer Science & Engineering, University of California San Diego, USA
[6]Center for Microbiome Innovation, University of California San Diego, USA

The human gut microbiome is estimated to harbor over 2 million unique protein-coding genes. Only a fraction of them is experimentally annotated and therefore require computational predictions. Community-wide experiments such as CAFA show that homology-based function annotation approaches are lacking and require more sophisticated approaches. We use protein families to predict residue-residue contacts and use them as constraints for the de novo structure predictions. The predictions are carried out using World Community Grid Microbiome Immunity Project (https://www.worldcommunitygrid.org/research/mip1/overview.do). Anyone can donate their spare computational time to the project. Until now, during 1.5 years of project duration, we were able to generate over 160,000 unique structural models, each representing a different gene family. Thus, effectively doubling the number of available protein structures. 3D structural models, instead of sequences, serve then as inputs for a deep learning-based function prediction method we developed. This approach enables us to achieve state-of-the-art accuracies in predicting gene ontology terms. We are now in a position to functionally annotate microbial genomes and metagenomes with higher coverage and accuracy. We may also start addressing microbe-microbe and host-microbiome protein-protein interactions to determine the mechanisms of microbiota-induced immune response.

Email: tkosciolek@ucsd.edu

Can bioinformatical methods surpass advanced experimental methods in recognition of amyloid protein aggregates ?

Malgorzata Kotulska[1], Michal Burdukiewicz[2], Natalia Szulc[1], Jakub W. Wojciechowski[1], Marlena Gąsior-Głogowska[1], Jaroslaw Chilimoniuk[3], Paweł Mackiewicz[3], and Vytautas Smirnovas[4]

[1]Wroclaw University of Science and Technology
[2]Warsaw University of Technology
[3]University of Wroclaw
[4]Vilnius University

The validation of new bioinformatic methods on a peptide data set is usually based on wet laboratory research. Validation experiments can include one of several different methods, such as spectroscopic studies, fluorescence methods or structural microscopy studies of AFM or EM. Microscopic methods are usually considered as the most accurate but they are very time consuming and expensive. Therefore other experimental methods are more often used to test the amyloid propensity of proteins. However, experimental methods may show discrepancies in their results and microscopy is not always decisive. We show how sensitive bioinformatic methods are to problematic or erroneous results of experiments and if they are capable of finding the source of such problems, even if they are trained on real and imperfect input data.

Email: Malgorzata.Kotulska@pwr.edu.pl

Multiple pathways of helix folding from kinetic coarse grained models

Krzysztof Kuczera[1,2] and Gouri Jas[3]

[1]Department of Chemistry, University of Kansas, Lawrence, KS 66045, USA
[2]Department of Molecular Biosciences, University of Kansas, Lawrence, KS 66045, USA
[3]Department of Pharmaceutical Chemistry University of Kansas, Lawrence, KS 66045, USA

Nanosecond laser temperature jump spectroscopy and molecular dynamics (MD) simulations were used to study the helix-coil transition in a 21-residue helix-forming peptide. Two relaxation times were detected a low pH, at which the single histidine in the peptide sequence is protonated. A slower component of 300-400 ns was assigned to helix-coil relaxation. A faster component of 20-35 ns was also characterized, indicating a complex path of peptide kinetics.

MD simulations were carried out to model the mechanism of formation of the peptide alpha-helical structure. A 12 microsecond MD trajectory in explicit solvent yielded structural and dynamic properties in good agreement with experimental data. Clustering and optimal dimensionality reduction were applied to produce low-dimensional coarse-grained models of the underlying kinetic network in terms of 2-5 metastable states. The high-entropy "coil" metastable set contains the largest number of structures, but the helix state was also structurally heterogeneous. The intermediate states contain the fewest structures, have lowest populations and the shortest lifetimes. As the number of considered metastable states increases, more intermediates and more folding paths appear in the coarse-grained models. One of these intermediates corresponds to the transition state for folding, which involves an "off-center" helical region over residues 11-16. The simulation data further suggest that the fast kinetic time scale should be assigned to correlated breaking/formation of blocks of several adjacent helical hydrogen bonds.

The same computational analysis was also applied to a 13-microsecond MD trajectory of the peptide with the neutral form of its histidine residue, corresponding to a higher pH . The loss of the histidine proton induces significant changes in the free energy landscape. This form has a higher helix content than the protonated peptide, in accord with experimental observations. Additionally, the kinetic network and folding pathway are markedly different.

Email: kkuczera@ku.edu

Quality assessment for intrinsic disorder predictions

Łukasz Kurgan[1]

[1]Virginia Commonwealth University, Richmond, US

Intrinsic disorder prediction is being pursued for over 30 years. Modern predictors rely heavily on machine learning and are widely used in computational and experimental studies. However, they lack quality assessment (QA) scores that quantify which residue-level predictions are more likely to be correct. The complicating factor is that the QA scores must be optimized for specific disorder predictors since these methods rely on different types of disorder annotations and have substantially different predictive architectures.

We will discuss first-of-its-kind toolbox of methods that provide accurate QA scores for ten popular disorder predictors. The QUARTER (QUality Assessment for pRotein inTrinsic disordEr pREdictions) tool relies on a machine learning model that is optimized for each of the ten disorder predictors, guided by an empirical feature selection and disorder predictor-specific putative propensities for disorder. Empirical tests on a large test dataset reveal that QUARTER generates high quality scores which significantly outperform the disorder propensities output by the original predictors. QUARTER is available as a convenient webserver at http://biomine.cs.vcu.edu/servers/QUARTER/.

The main application of the QA scores produced by QUARTER is annotation/selection of a high-quality subset of residue-level disorder predictions. We show that when combining results from the ten disorder predictors, the QA scores can be used to identify 40% of residues for which disorder is predicted with 95% precision.

Email: lkurgan@vcu.edu

Sig2Lead: Combining pharmacogenomics and cheminformatics towards enhanced lead compound identification

Jaroslaw Meller[1]

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

TBA

Email: mellerj@ucmail.uc.edu

Feature engineering for biological data

Robert Nowak[1]

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

The results obtained from machine learning models depends on quality of the training data. The crucial aspect to obtain acceptable results is properly developing the attribute set. The number of attributes should be small enough to prevent over-fitting. It is big problem in models created for large biological sequence analysis. The deep learning techniques can not help in this case, because there is no large enough labelled training datasets available.

In presented work We are focused on feature engineering methods. Feature engineering involves use of external sources of data. Such methods require human expert knowledge of the problem, but the generated attributes can be better correlated, therefore results could be significantly better. For example the decision, if customer is willing to take a taxi, could be better correlated with weather conditions then the geographical coordinates.

We present two real cases, where we used several models (naive bayesian, random forest, logistic regression, gradient boosting, neural network) for customer decision prediction in motor insurance market. The feature engineering were crucial to obtain significantly better results of classification, and consequently to convince the company management to use our models in production. We discuss such methods for biological sequence analysis.

Email: r.m.nowak@elka.pw.edu.pl

Predictive models to decode gene regulatory regions

Uwe Ohler[1]

[1]Max-Delbrück-Centrum für Molekulare Medizin, Berlin, Germany

TBA

Email: uwe.ohler@mdc-berlin.de

Bioinformatics discovery of novel surprising pseudokinase families

Marcin Gradowski[1] and Krzysztof Pawłowski[1]

[1]Warsaw University of Life Sciences SGGW

Protein kinases are an 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 such novel families.

Some kinases lack certain elements of their active sites and are thus believed to be enzymatically inactive, hence the term "pseudokinases". Surprisingly, as our collaborators, dr Tagliabracci and colleagues were able to show, some of these actually perform other enzymatic activities utilizing modified active sites.

The SELO family that we discovered previously as unusually well-conserved group of putative kinases, although having a protein kinase-like three-dimensional structure, utilizes ATP in a manner different from all other kinases, for AMPylation of protein substrates (Cell, 2018). We also show that this function is conserved in SELO in evolution (bacteria, yeast, humans) and important for response to oxidative stress.

SidJ, an effector from Legionella, is one of many effectors delivered into the host cell, to massively rewire the signalling pathways. We predicted it to possess a kinase-like fold with a strangely modified active site. Indeed, crystal structure revealed a modified kinase-like structure with a double active site. However, functional experiments showed that SidJ is actually a polyglutamylase (Science, 2019).

We show how bioinformatics, hand in hand with structural biology, biochemistry and cell biology, can drive novel research projects. We argue that search for distant members of established enzyme families can bring discoveries such as variations of known functions or novel, unexpected functions.

Email: krzysztof_pawlowski@sggw.pl

Genomic Features: An Actionable Key for Gut Microbiota Modulation in Cancer Treatment

Agata Szymanek[1]

[1]Ardigen S.A.

Building robust machine learning models

Aneta Polewko-Klim[1], Krzysztof Mnich[2], Wojciech Lesiński[1], Radosław Piliszek[2], Bogumił Sapiński[3], and Witold Rudnicki 1 2[3]

[1]Institute of Informatics, University of Białystok,
[2]Computational Centre, University of Białystok,
[3]Interdisciplinary Centre for Mathematical and Computational Modelling,

Background: Modern experimental techniques deliver data sets containing profiles of tens of thousands of potential molecular and genetic markers that can be used to improve medical diagnostics. We propose methodology arise due to limited sample size and feature selection. It is based on comprehensive cross-validation protocol, that includes feature selection within cross-validation loop and classification using machine learning. This methodology allows for estimation of biasaes inherent in the machine learning. Protocol was utilised for building several models based on sets of variables of varying sizes that were selected using three different feature selection methods.

Results: The significant biases due to feature selection procedure and split of the sample between training and validation sets were observed. The size of the bias depends on the size of experimental sample. Good correlation between performance of the models in the internal and external cross-validation was observed, confirming the robustness of the proposed protocol and results.

Conclusions: We have developed a protocol for building predictive machine learning models. The protocol can provide robust estimates of the model performance on unseen data. It is particularly well-suited for small data sets. We have applied this protocol to develop prognostic models for neuroblastoma, using data on copy number variation and gene expression.

Email: W.Rudnicki@icm.edu.pl

Integration of omics data in a pathway related context to study complex disease aetiology

Ugur Sezerman[1]

[1]Acibadem University, Istanbul, Turkey

TBA

Email: ugur.sezerman@acibadem.edu.tr

Combining Big Data and Machine Learning for predicting ribonucleic acid-ligand interactions

Filip Stefaniak[1], Pietro Boccaletto[1], and Janusz M. Bujnicki[1]

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

Currently, most of the registered drugs are small molecule compounds. Computational methods play a pivotal role in the early stages of drug discovery and are widely applied in virtual screening, structure optimization, and compound activity profiling. Over the last decades, almost all the attention in medicinal chemistry has been directed to protein-ligand binding, and computational tools have been created with such targets in mind. However, with growing discoveries of functional RNAs and their possible applications, RNA have gained considerable attention as possible drug targets. This flow of discovery was followed by adapting existing protein-based computational tools for RNA applications, as well as active development of new RNA-tailored methods. However, due to the difference in nature of RNA from that of proteins, especially its tendency to use morphological plasticity (conformational change in ligand binding), the modeling of RNA remains a challenging task.

We will present two new computational tools for predicting RNA-ligand interactions. One is a semi-automatic modeling procedure, which involves input data pre-processing (RNA and ligand files), molecular docking, re-scoring of models of complexes, clustering of poses and visualization of results. The second tool that we will present is a new scoring function, AnnapuRNA, for scoring RNA-ligand complexes obtained, e.g., from molecular docking. It is based on statistical data derived from the experimentally solved RNA-ligand complex structures and predictive models obtained using machine learning techniques. According to our benchmarking tests, AnnapuRNA outperforms other tested scoring functions, including rDock and our previous method LigandRNA.

Email: `fstefaniak@genesilico.pl`

ML@Google for Life Science

Szyman Wąsik[1]

[1]Google

The recent development of the machine learning methods gives a lot of benefits in many research areas but also creates a lot of challenges. In this talk I would like to present how Google is trying to help community to solve the most important of them by providing dedicated software, services and hardware. This includes product such as AI Platform, TensorFlow or Tensor Processing Units. I will also cover the most successfull applications of these products in the area of life sciences, such as diagnosing diabetic retinopathy, detecting cancer metastases, predicting tasks in healthcare based on electronic health records, predicting properties of molecules or understanding genomic data. All these applications were successfully tackled by Google researchers, using the recent advancement in machine learning, primally in deep learning.

Email: szymon.wasik@cs.put.poznan.pl

Machine learning in multi-omics medicine

Maciej Wiznerowicz[1]

[1]Poznan University of Medical Sciences, University Hospital of Lord's Transfiguration, Medical Oncology; International Institute for Molecular Oncology, Poznan, Poland

Cancer progression involves the gradual loss of a differentiated phenotype and acquisition of progenitor and stem cell-like features. Here, we provide new stemness indices for assessing the degree of oncogenic dedifferentiation. We took advantage of an innovative one-class logistic regression machine learning algorithm (OCLR) to extract transcriptomic and epigenetic feature sets derived from non-transformed pluripotent stem cells and their differentiated progenies. Using OCLR, we were able to sort TCGA tumor samples by stemness phenotype and identify previously undiscovered biological mechanisms associated with the dedifferentiated oncogenic state. Analyses of tumor microenvironment revealed the correlation of cancer stemness with immune checkpoint expression and infiltrating immune system cells not previously anticipated. We have shown the de-differentiated oncogenic phenotype increased in the metastatic tumor that further justify their more aggressive phenotype. Application of our stemness indices reveals features of intra-tumor heterogeneity in molecular profiles obtained from the single-cell analyses. Finally, the machine learning-based indices allowed for the identification of chemical compounds and novel targets for the cancer therapies aiming at tumor differentiation. Our findings provide new prognostic signatures that enable cancer biologists and oncologists to quantify the impact of tumor stemness on outcome across cancer types and may help to pave the way for progress in treatment strategies for cancer patients.

Email: maciej.wiznerowicz@gmail.com

General quality assessment of 3D RNA structures using machine learning techniques

Marcin Zabłocki[1], Marta Szachniuk[2,1], and Maciej Antczak[1,2]

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

General quality assessment of 3D RNA models generated by in silico prediction methods is crucial to identify native or near-native 3D RNA structures. Nowadays, there are many, freely available computational methods for RNA 3D structure prediction, however without the reference, experimentally determined RNA 3D structure it is difficult to reliably rank them in terms of practical usefullness. Here, we propose applying machine learning methods, involving traditional and deep learning techniques to infer the relationships commonly observed in 3D RNA structures allowing us to reliably predict the quality of RNA 3D models. We analyse experimentally determined RNA 3D structures stored in non-redundant RNA 3D structures repository. Due to scarcity of the experimentally determined RNA 3D structures we are currently generating RNA 3D models using all state-of-the-art, computational methods for RNA 3D structure prediction. Thus, machine learning-oriented methods are able to harness the wide range of 3D conformational space. We also solved the problem of variable-length RNA 3D structure representation crucial for proper application of machine learning techniques by snapshotting local 3D motifs and applying an ensemble of local quality scores to assess the whole 3D RNA structure quality. Novel general quality assessment approach that does not need the reference RNA 3D structure to reliably rank RNA 3D models will be breakthrough in the field of RNA structural bioinformatics.

Email: m.zablo+bit2019@gmail.com

Posters

Prediction of Mg-RNA interactions : Statistical potential approach

Masoud Amiri Farsani[1], Michał Boniecki[1], Pritha Ghosh[1], Filip Stefaniak[1], and Janusz M. Bujnicki[1]

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

Folding RNA in the presence of ligands is an essential subject that is very interesting for scientists. Biological molecules are so complex that their interactions cannot be quickly inferred from simple mathematical equations derived from first principles. To overcome this problem, we have resorted to a knowledge-based approach. SimRNA is a method that has been developed by our group. SimRNA uses a method for RNA 3D structure prediction and folding simulations. Currently, we are developing SimRNA-L, a new method that enables us to fold RNA in the presence of ligands. Recently, we developed a prototypical method for modeling the interactions of RNA with Mg2+ ions, which are among the simplest and the most commonly observed ligand in RNA structures. Now total energy of the system has three parts, 1- RNA energy, 2- RNA- Mg2+interaction and 3- Mg2+- Mg2+ interaction. Sampling states and considering energy landscape and also clustering results will allow us to find structures which are close to nature. Our new method successfully predicts preferred Mg2+-binding sites in RNA structures. We are currently extending this approach to bigger organic ligands, which will help us in the development of a general purpose method for prediction of RNA-ligand 3D structures.

Email: mfarsani@genesilico.pl

Gene cooccurrence analysis for prediction of gene function

Bartosz Baranowski[1] and Krzysztof Pawłowski[2]

[1]Institute of Biochemistry and Biophysics Polish Academy of Sciences,
[2]Warsaw University of Life Sciences Department of Experimental Design and Bioinformatics

Many genes in microbial genes remain functionally uncharacterised. Understanding signalling and metabolic pathways involving such genes is essential for deeper understanding of microbial biology, and also mechanisms of infectious diseases. It is well-known that genes which co-occur across genomes are more likely to share similar biological functions than random pairs of genes. This has been exploited in some bioinformatics tools for functional relationship prediction, e.g. STRING, Prolinks. This gene co-occurrence approach suffers from uneven sampling of the microbial world by genome sequencing projects, e.g. there are thousands of available genomes of different Escherichia coli strains. To remedy this, we propose a novel algorithm that allows collapsing the co-occurrence relationships at different taxonomic levels, e.g. strains or species, and limiting the analysis to selected taxonomic groups. A rigorous statistical assessment with provision for multiple testing is provided. Thus, sets of significantly co-occurring genes can be elucidated for query genes. The method is implemented in a prototype server. Usage examples are provided for a number of Legionella pneumophila uncharacterized effector proteins known for their involvement in infection.

Email: bartosz.piotr.baranowski@gmail.com

Reduction of a helical bias during derivation of statistical potential for SimRNA

Michał Boniecki[1], Grzegorz Łach[2], and Janusz Bujnicki[1,3]

[1]International Institute of Molecular and Cell Biology in Warsaw
[2]Faculty of Physics, University of Warsaw
[3]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University

The molecules of the ribonucleic acid (RNA) perform a variety of vital roles in all living cells. We have developed a computational method for molecular simulations of RNA, especially for perdition 3D structure, named SimRNA.

SimRNA is based on a coarse-grained representation of a nucleotide chain, a statistically derived energy function, and Monte Carlo methods for sampling of the conformational space. The backbone of RNA chain is represented by P and C4' atoms, whereas nucleotide bases are represented by three atoms: N1-C2-C4 for pyrimidines and N9-C2-C6 for purines. All base-base interactions were modeled using discrete three-dimensional grids built on local systems of coordinates.

All terms of the energy function used were derived from a manually curated database of crystal RNA structures, as a statistical potential. The main idea of deriving statistical potential is that: we rely on what we see. In SimRNA, the energy function is decomposed into local terms (controlling backbone conformation), and base-base interaction terms that can be divided into stacking and egde-edge contributions. We have noticed that both local terms and stacking terms posses deep local minima corresponding to the helical conformation. In fact, they introduce helical bias, which can be seen in the testing runs. From one hand the helical bias makes the method stable, on the other hands it harms recapitulation of peculiarities, motifs of irregular trace of a backbone.

We made attempts of re-balancing of the input database, which is used for derivation the statistical energy terms. Using predefined classifier, we identified regions corresponding to the helical conformations and omitted them during derivation of the statistical terms. We re-normalized the terms accordingly.

Email: mboni@genesilico.pl

Improving scoring of protein-peptide models with machine learning: looking for relevant similarity measures

Maciej P. Ciemny[1,2] and Sebastian Kmiecik[1]

[1]Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Warsaw, Poland
[2]Faculty of Physics, University of Warsaw, Warsaw, Poland

The modeling of protein-peptide complexes is a challenging problem[1]. The peptides are highly flexible molecules and their binding to proteins is governed by transient, non-specific interactions. In many cases, the docking methods produce near-native models but fail to correctly identify them with scoring functions. This is the case of well-established CABS-dock method[2,3] (`http://biocomp.chem.uw.edu.pl/CABSdock/`) for flexible docking of peptides to proteins. CABS-dock uses structural clustering using similarity measure based on root mean squared deviation (RMSD) of atomic positions. Unfortunately, RMSD is not the perfect measure of the models similarity or proximity to the correct native-like structure. To counter this problem, we have developed a set of similarity measures based on protein-peptide contact patterns. Here, we present the performance of the measures on a standard protein-peptide docking benchmark set and their applications to modeling cases using the CABS-dock method. The measures can improve existing and lead to new, machine-learning scoring methods.

1. Ciemny, M. et al. Protein–peptide docking: opportunities and challenges. Drug Discovery Today 23, 1530–1537 (2018). 2. Kurcinski, M. et al. CABS-dock standalone: a toolbox for flexible protein-peptide docking. Bioinformatics (2019). doi:10.1093/bioinformatics/btz185 3. Kurcinski, M., Jamroz, M., Blaszczyk, M., Kolinski, A. & Kmiecik, S. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. Nucleic Acids Res. 43, W419–24 (2015).

Email: `maciej.ciemny@gmail.com`

Gene expression profile of BRAF-positive and BRAF-negative papillary thyroid carcinomas and impact of co-existing TERTp mutation.

Dagmara Rusinek[1], Marta Cieślicka[1], Jolanta Krajewska[1], Aleksandra Pfeifer[1], Tomasz Tyszkiewicz[1], Sylwia Szpak-Ulczok[1], Małgorzata Kowalska[1], Jadwiga Żebracka-Gala[1], Monika Halczok[1], Ewa Chmielik[3], Ewa Zembala-Nożyńska[3], Agnieszka Czarniecka[2], Barbara Jarząb[1], and Małgorzata Oczko-Wojciechowska[1]

[1]Maria Skłodowska-Curie Institute - Oncology Center, Gliwice Branch, The Department of Nuclear Medicine and Endocrine Oncology,
[2]Maria Skłodowska-Curie Institute - Oncology Center, Gliwice Branch, The Oncologic and Reconstructive Surgery Clinic,
[3]Maria Skłodowska-Curie Institute - Oncology Center, Gliwice Branch, Tumor Pathology Department,

BRAFV600E mutation is the most common genetic alteration in papillary thyroid carcinoma (PTC), however, its prognostic significance is controversial. One of the less common alterations in PTC is mutation of TERT promoter (TERTp) which alone is responsible for greater aggressiveness of the disease. There are evidences that co-existence of those two mutations have negative impact on clinical outcomes in patients with PTC. The aim of our study was the search for potential differences in PTCs gene expression profiles depending on mutation status in BRAF and TERTp.

The PTC cohort in our study consisted of 54 cases [16 being BRAF(+) and TERTp(-), 8 BRAF and TERTp(+) and 30 PTCs without detected BRAF and TERTp mutations]. Affymetrix Human Gene 1.0 ST arrays were used. Microarray data has been pre-processed using the fRMA and ComBat algorithms. The Student's t-test and FDR were used in the analysis of genes expression. PTCs with BRAFV600E, with TERTp mutation and without these genetic alterations were compared to each other for potential differences in gene expression profile. Additionally, BRAF-like, RAS-like score (BRS) and Thyroid Differentiation Score (TDS) were calculated.

In comparison between BRAF(+) and BRAF(-) samples over 2500 genes showed statistically significant difference in expression level (FDR$<$0.05). In comparison between BRAF(+) TERTp(-) and BRAF(+) TERTp(+) only 9 genes expressions were statistically significant. TDS and BRS were affected by BRAF mutation status, but not TERTp mutation.

Obtained results do not show major changes in gene expression profile in TERTp(+) and BRAF(+) PTC's comparing to BRAF(+) ones. Lack of cases with the presence of only TERTp mutation significantly impairs our analysis of these mutations impact. Further studies are needed.

Email: marta.cieslicka@io.gliwice.pl

Ligand Dissociation Pathways in T4 Lysozyme L99A

Sylwia Czach[1,2], Aleksander Oskroba[1], and Jakub Rydzewski[1]

[1]Institute of Physics, Nicolaus Copernicus University, Grudziadzka 5, 87-100, Torun, Poland
[2]Department of Biochemistry, Nicolaus Copernicus University, Lwowska 1, 87-100, Torun, Poland

Recent developments in enhanced sampling methods showed that it is possible to find ligand unbinding pathways with spatial and temporal resolution inaccessible to experiments. Such techniques should provide an atomistic definition of possibly many reaction pathways, otherwise it may lead either to overestimating energy barriers, or inability to sample hidden energy barriers that are not captured by simple reaction pathway estimates. We provide an official Plumed 2 module that implements a method which is able to sample the reaction pathways of the ligand-protein dissociation process. The method is based on enhanced sampling and non-convex optimization methods. The module, called MAZE, requires only a crystallographic structure to start a simulation, and does not depend on many ad hoc parameters. To present its applicability and flexibility, we provide several examples of ligand unbinding pathways along transient protein tunnels reconstructed in a model ligand-protein system.

Email: `jr@fizyka.umk.pl`

Initial analysis of data obtained to identify potential markers of Chronic Obstructive Pulmonary Disease

Bartłomiej Deszcz[1] and Krzysztof Pawłowski[1]

[1]Warsaw University of Life Sciences

Chronic Obstructive Pulmonary Disease is currently one of the most common causes of death. The main symptom of COPD is pulmonary damage, that reduces lung function and increased susceptibility to infection. Currently, the only classification used for this disease, known as GOLD, is based on spirometric measurements with the assessment of current symptoms and the risk of exacerbations. The lack of specific symptoms and, because of that, the difficulty in precisely diagnosing the disease, results in late detection and less effective treatment. Therefore, it is necessary to identify biomarkers that enable early diagnosis COPD symptoms. Accordingly, we decided to conduct long-term studies of healthy people and patients diagnosed with COPD. For this purpose, data from the lifestyle/symptoms questionnaire, clinical chemistry/blood samples, proteomics and spirometry were collected, which were then subjected to bioinformatics analyses. Here we present an analysis of data obtained in the process of routine visits for 70 subjects involved in the project of identifying potential COPD markers . We investigated the relationships between patients' subjective assessment of their health based on the responses in the questionnaire and the objective determination of the stage of COPD, which involved medical examination and the results of spirometry. We also checked whether the answers in the questionnaire depended on the gender, and what were the most frequently reported symptoms. The obtained results allowed to assess the usefulness of the health/lifestyle questionnaire in the process of diagnosing COPD and to check how the stages of the disease affect the respondents' answers. Further, we explored the relationships between proteomics and clinical chemistry data and disease stages.

Email: bartek_deszcz@wp.pl

Hierarchical division approach to pan-genome analysis.

Paulina Dziadkiewicz[1] and Norbert Dojer[2]

[1]Warsaw University of Technology,
[2]University of Warsaw

The constant growth of genomic data intensifies development of data structures and algorithms to analyse them efficiently. The classical way of multiple sequence comparative analysis is their alignment. The requirement of more advanced techniques to store, process and visualize multiple sequences leads to arising of a new research field called 'pan-genomics'. This work is focused on providing a tool for building a graph model of given multiple genome alignment and discover relations between aligned sequences. A new structure to represent the relationships - called Consensus Tree - is proposed. Each node of this tree has a subset of aligned sequences assigned and every non-leaf node has at least two children nodes that form a partition of the sequences assigned to their parent into more homogeneous subsets. Moreover, an averaged consensus sequence is assigned to every node. The results of our approach may be inspected in our interactive visualization tool. We demonstrate its functionality using the pan-genome of Ebola virus.

Email: pedziadkiewicz@gmail.com

Identification and profiling microbial signatures across various urban biomes

Dedan Githae[1], Paweł Łabaj[1], Wojciech Branicki[1], Gabriella Mason-Buck[2], Alexandra Graf[3], Josef Moser[3], Przemysław Kiljan[1], Michał Kowalski[1], and Witek Wydmanski[4]

[1]Jagiellonian University, PL
[2]King's College London , UK
[3]FH Campus Wien, AT
[4]AGH University of Science and Technology, PL

Microbial communities are abundant in our environs but remain unknown due to their complexity. Microbial diversity was studied via 16s rRNA sequencing, but emergence of high-throughput NGS technologies availed entire metagenomes from environmental samples. In addition, advanced machine learning algorithms allow exploration of whole metagenomic sequences. Establishing a genomic baseline across metagenomic sources can be exploited by using datasets of known geographical origin to supervise a machine learning model in learning to recognize the source of the sample, as well as other important characteristics such as level of urbanization or density of population. Aim: To develop machine-learning approach to identify unique microbial profiles across cities.

Email: dedangithae@yahoo.com

Assemblying differential transcripts with overlap graphs

Julia Herman-Izycka[1] and Bartek Wilczynski[1]

[1]Institute of Informatics, University of Warsaw

Metatranscriptomics is an increasingly popular approach to investigate microbial communities. Paired with differential expression analysis provides way to analyse changes in both composition and activity of bacteria in response to the change of environment. In metatranscriptomic studies it is usually not known which species are present in the samples, and reference genomes of many of those species are not sequenced yet. Therefore assembly paired with read mapping, counting and differential expression analysis tools is a standard workflow for that kind of experiments.

We consider usage of de Bruijn graph and propose using an overlap graph to directly assembly transcripts with different abundance between conditions (i.e. fold-change above some threshold). We propose a few simple heuristics to transcript assembly overlap graph and investigate ability to use those on experimental data from metatranscriptomic study of mice feacal microbiota change in response to high-fat diet.

Email: `j.herman-izycka@mimuw.edu.pl`

PARNASSUS based encoding to find better homologs for RNA sequence search

Dharm Jain 1[2], Chandran Nithin[1], and Janusz Bujnicki 1[3]

[1]International Institute of Molecular and Cell Biology, Warsaw, Poland
[2]Warsaw University of Technology, Warsaw, Poland
[3]Adam Mickiewicz University, Poznan, Poland

Sequence searches have always been an intriguing problem in the domain of bioinformatics, with applications ranging from forming phylogenetic trees to predicting structure-function relationships for homologs, often influencing related research in various aspects. Numerous highly optimized and accurate tools have been developed for related applications in the analysis of proteins. Primarily, these methods use the protein sequence information to perform searches in biological databases. Over the years, similar tools have also been developed for RNAs. BLASTn is such a method which uses the primary sequence of RNAs represented by 4 letters. BlastR in contrary converts the input from a 4 letter to a 16 letter representation. In this work, we present Protein like Alphabet for RNA Secondary Structure Unified with Sequence (PARNASSUS) which explores the idea of representing RNA primary sequences combined with secondary structure into a single sequence using a 20 letter code similar to proteins. Such an approach allows us to tweak and use tools developed for proteins to process datasets of RNAs, for instance, an iterative approach like PSI-Blast. An application for such a method could be to build accurate clusters of sequence families in an automated manner. As an example, the customized BLASTp and PSI-BLAST tools detect a high degree of similarity between the Downstream-peptide RNA and Glutamine Riboswitch. These two families are not classified in a single clan as per Rfam but are reported to have similar structures and functions. PARNASSUS is a very promising tool, that will allow highlighting not known till now connections between various RNA families, and should outperform currently available tools used for RNA sequence searches.

Email: dsjain@genesilico.pl

LCR-BLAST: Adaptation of BLAST to search for similar low complexity regions

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

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

Low complexity regions (LCRs) in proteins occur significantly more often than by chance. About 14% of proteins contain these regions. They often play key roles in protein functions, structure and properties. The abundance of LCRs and their striking similarity to each other led us to the hypothesis that we can predict functions of LCRs by searching for similar ones where at least some would have a known function. However, currently available methods for searching for similar protein sequences are focused on searching for similar evolutionary relationships but in high complexity regions. These methods mask LCRs during similarity search. In order to overcome this limitation we adopted one of the most popular tool to search for similar LCRs.

We have many modifications of the BLAST method, e.g. PSI-BLAST, PHI-BLAST or DELTA-BLAST. Here we present the LCR-BLAST which effectively improves the search for similar LCRs. Proposed modifications are: change of some default parameters, implementation of identity scoring matrix and introduction of new statistics to assess similarity between proteins. Changes to the default parameters are mainly related to LCRs masking. New statistics helps the user to compare LCRs with pattern features and is length independent.

Email: patryk.jarnot@polsl.pl

Differences in genomic structure between two cattle breeds

Barbara Kosinska-Selbi[1], Magdalena Frąszczak[1], Tomasz Suchocki 1[3], Christa Egger-Danner[2], Hermann Schwarzenbacher[2], and Joanna Szyda 1[3]

[1]Biostatistics group, Wroclaw University of Environmental and Life Sciences, Kozuchowska 7, 51-631 Wroclaw, Poland
[2]ZuchtData EDV-Dienstleistungen GmbH, Dresdner Straße 89/19, 1200 Vienna, Austria
[3]National Research Institute of Animal Production, Krakowska 1, 32-083 Balice, Poland

In this study we investigate differences in linkage disequilibrium (LD) structure of two Austrian cattle breeds – Fleckvieh and Braunvieh. The data set comprised of 2 984 cows (1 999 – Fleckvieh and 985 Braunvieh) genotyped using the Geneseek Genomic Profiler HD BeadChip, consisting of 76 932 single nucleotide polymorphisms (SNPs). SNPs with the call rate below 95% and a minor allele frequency below 5% were removed from final analysis LD between pairs of linked SNPs was quantified with the r2 statistic using Beagle 4.1, what resulted in 104 503 568 estimates for Fleckvieh and 105 401 852 estimates for Braunvieh. The estimates for r2 were used to generate two covariance matrices in order to significant differences between two breeds. Both LD matrices were decomposed using principal components analysis for non-overlapping windows containing 50 SNPs. For each such window, functions of the first two principal components were compared between breeds in order to quantify the variability between breeds. Results shown that in general, both breeds showed a similar structure of genetic variability, but some differences appeared on BTA19, BTA17, BTA16, BTA15, BTA14, BTA13 and BTA12. We also used a method which allows to measure changes in shape and orientation of the matrix based on comparing two covariance matrices similar size. Based on information obtained from principal component analysis, we used statistics which allowing for the assessment of matrix differences. First statistics measure the contribution between matrix differences in orientation, second measure changes in the shape of the matrix. Changes in variability of covariance matrix were found for BTA4, BTA5, BTA7, BTA13, BTA14, BTA17, BTA24 and BTA29.

Email: barbara.kosinska@upwr.edu.pl

Comparison of kNN and k-means optimization methods of reference set selection for improved CNV callers performance

Wiktor Kuśmirek[1], Agnieszka Szmurło[1], Marek Wiewiórka[1], Robert Nowak[1], and Tomasz Gambin[1]

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

There are over 25 tools dedicated for the detection of Copy Number Variants (CNVs) using Whole Exome Sequencing (WES) data based on read depth analysis.

The tools reported consist of several steps, including: (i) calculation of read depth for each sequencing target, (ii) normalization, (iii) segmentation and (iv) actual CNV calling. The essential aspect of the entire process is the normalization stage, in which systematic errors and biases are removed and the reference sample set is used to increase the signal-to-noise ratio.

Although some CNV calling tools use dedicated algorithms to obtain the optimal reference sample set, most of the advanced CNV callers do not include this feature.

We used WES data from the 1000 Genomes project to evaluate the impact of various methods of reference sample set selection on CNV calling performance of three chosen state-of-the-art tools: CODEX, CNVkit and exomeCopy. Two naive solutions (all samples as reference set and random selection) as well as two clustering methods (k-means and k nearest neighbours (kNN) with a variable number of clusters or group sizes) have been evaluated to discover the best performing sample selection method.

The performed experiments have shown that the appropriate selection of the reference sample set may greatly improve the CNV detection rate. In particular, we found that smart reduction of reference sample size may significantly increase the algorithms' precision while having negligible negative effect on sensitivity. We observed that a complete CNV calling process with the k-means algorithm as the selection method has significantly better time complexity than kNN-based solution.

Email: kusmirekwiktor@gmail.com

Measurement biases affecting RNA sequencing and methods of its normalization.

Paweł Kuś[1], Roman Jaksik[1], and Marek Kimmel[1,2]

[1]Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
[2]Department of Statistics, Rice University, Houston, TX USA

Rapid development of RNA sequencing methods has made them the leading approach used for transcriptome-level studies. Most experimental strategies require complicated material preparation process, whose stages, especially cDNA synthesis and PCR amplification, are reported to be sensitive features associated with the gene structure. These features include gene length and nucleotide composition, which affect sequencing results and may bias subsequent analysis, such as identification of molecular markers specific for different cell types.

This work investigates the impact of 1) factors affecting results of RNA sequencing and 2) existing normalization methods that take some of those factors into account. We obtained RNA sequencing data with the synthetic spike-in material added in known concentrations from NCBI database and performed standard data processing procedure using STAR and bedtools coverage programs to map and count reads. Then FastQC program and set of methods from EDAseq and RSeQC packages were used to assess the influence of selected factors: GC-content, gene length, sequence duplication level and RNA-degradation on measured level of expression.

Among factors examined we identified GC content-related effect as the dominant source of bias and applied appropriate methods from EDAseq package to minimize its effect. We show that transcript GC content significantly affects coverage levels in a way that differs between the analyzed samples. This indicates that the correction based on nucleotide composition is a necessity, however currently available methods not always address this problem properly, increasing the need for the development of novel data normalization strategies.

The work is financed by Polish National Science Centre grant 2016/23/D/ST7/03665.

Email: pawel.kus@polsl.pl

To "see" or not to "see"? Is AI able to detect external world?

Wiktor Łachmański[1], Jakub Rydzewski[1], and Wiesław Nowak[1]

[1]1 Institute of Physics, Faculty of Physics, Astronomy and Informatics, Nicolaus Copernicus University, 87-100 Torun, Poland

Optogenetics is a relatively new research method allowing for studies of neurons using light. Genetically modified neurons may be externally activated or deactivated using photons. Biological activity, similarly to a normal vision process, is triggered by photon-sensitive proteins channelrhodopsins. A typical channelrhodopsin [1] (ChR) is a transmembrane ion channel, consistsing of seven alpha-helices and a retinal moiety which undergoes photoisomerization from all-trans to 13-cis form. This local kick leads to a cascade of events in protein structure that result in opening of this ion channel. In order to develop new tools for light-activated protein studies, we have performed molecular dynamics (MD) simulations of ChR2, imitating a continous mode light excitation of the retinal. The protocol mimics flip-flop type switching between trans-cis retinal states in 1 ns intervals during 100 ns trajectory. We applied a several artificial intelligence (AI) based clustering algorithms to analyze such a trajectory, in part using reduced conformational space representations. A long standing goal of our efforts is to check whether one can use AI to recognize effecively the state of similar protein systems (ON/OFF). This may help to determine structural determinants of channelrhodopsins' variants activities.

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

[1] H. E. Kato et al. Crystal structure of the channelrhodopsin light-gated cation channel. Nature 482, 369-374 (2012)

Email: vicolls@fizyka.umk.pl

DeepCoil - a fast and accurate prediction of coiled-coil domains in protein sequences

Jan Ludwiczak 1[2], Aleksander Wiński[1], Krzysztof Szczepaniak[1], Vikram Alva[3], and Stanisław Dunin-Horkawicz[1]

[1]Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland
[2]Laboratory of Bioinformatics, Nencki Institute of Experimental Biology, Pasteura 3, 02-093 Warsaw, Poland
[3]Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tuebingen 72076, Germany

Coiled-coils domains are present in approximately 15% of proteins and they are involved in a plethora of biological functions such as signal transduction, molecular transport and mediation of oligomerization processes [1]. Thus, their reliable annotation is crucial for studies of protein structure and function. Here, we report DeepCoil [2], a novel neural network-based tool for the detection and localization of coiled-coil domains in protein sequences. DeepCoil predictions are based either on a sequence information alone (DeepCoil_SEQ) or on a sequence and a profile derived from homologous sequences (DeepCoil_PSSM). In a rigorous benchmark both DeepCoil variants outperformed current state-of-the-art methods and detected many coiled coils that remained undetected by other methods. This higher sensitivity of DeepCoil opens up a possibility of an accurate, genome-wide annotation of coiled-coil domains without the time-consuming profile generation. We will also present strategies for improvement of predictor performance through large-scale distributed optimization of hyper-parameters and network architecture.

DeepCoil is freely available as a web server at https://toolkit.tuebingen.mpg.de/#/tools/deepcoil and as a standalone tool at https://github.com/labstructbioinf/DeepCoil.

[1] Lupas AN, Bassler J, Dunin-Horkawicz S. The Structure and Topology of *alpha*-Helical Coiled Coils. Subcell Biochem., 2017, 82:95-129

[2] Ludwiczak J, Wiński A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil - a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics., 2019, Jan 2

Email: `j.ludwiczak@cent.uw.edu.pl`

The PDBrt: a free database of drug - target residence time

Magdalena Ługowska[1] and Marcin Pacholczyk[1]

[1]Institute of Automatic Control Silesian University of Technology Akademicka 16 44-100 Gliwice Poland

The residence time of a drug in its molecular target is becoming a key parameter in the design and optimization of new drugs, as it can reliably predict drug efficacy in vivo. In 2006 Copeland introduced the drug-target residence time concept which dictates a significant proportion of pharmacological activity in vivo [1]. The traditional in vitro methods perceive drug-target interactions only in terms of the affinity in equilibrium; the residence time concept takes into account the conformational dynamics of the target molecules that affect the binding and dissociation of the drug [2]. Experimental approaches to binding kinetics and target ligand complex solutions are currently available, but known bioinformatics databases do not usually store information about the drug residence time in its molecular target. The Protein Data Bank Residence Time (PDBrt) is a free, non-commercial repository for 3D protein-ligand complex data with measured drug residence time inside binding pocket of the specific biological macromolecules deposited in the Protein Data Bank [3]. The PDBrt is implemented using Python/DRF/HTML/CSS and contains information about both the protein and the drug separately as well as the entire complex and time of drug residence inside the protein. Collected dataset consists of ca. 150 crystallographic structures of protein-ligand complexes with known drug-target residence time (measured using experimental methods) and can be crucial for many computational or machine learning studies on drug binding/unbinding in biological systems. The work was supported by grant No. 02/010/BK_18/0102 from Silesian University of Technology, Gliwice, Poland and co-financed by the European Union through the European Social Fund (grant POWR.03.02.00-00-I029).

1.Copeland R. et al. Nat Rev Drug Discov.2006 Sep;5(9):730-9 2.Copeland R.2016 Feb;15(2):87-95 3.Berman H.M. et al.2000. 28: 235-242

Email: magdalena.lugowska@polsl.pl

Classification of Alzheimer's Disease patients based on whole-genome sequencing data – challenges and limitations

Marlena Osipowicz[1], Bartek Wilczynski[1], and Magdalena A. Machnicka[1]

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

Despite huge amount of data from genome-wide association studies (GWAS) and whole-genome sequencing (WGS) the genetic background of Alzheimer's disease (AD), which is partially heritable, is not fully understood yet. Machine learning approaches applied to genotype-based classification of AD patients and healthy controls using GWAS data have reported classification accuracies of 0.65-0.9. However, since the estimated heritability of AD is 0.6-0.8, higher classification accuracies may result from overfitting.

We have applied feature selection and classification to single-nucleotide polymorphisms (SNPs) from WGS and GWAS data from two patient groups. Classification performance was around 0.98 (AUC) when feature selection was performed on the complete dataset and much lower (AUC 0.6)when it was done on the training set only. Application of models trained on one dataset to classification of a different dataset showed that the information learned by the models is mostly shared between datasets.

Our results suggest that application of classification algorithms to AD genome-wide datasets is prone to overfitting if feature selection is performed before division of data into the training and test set. The expected classifier performance is between 0.55 and 0.7 (AUC) for currently available dataset sizes. Our results are in agreement with the estimated AD heritability, especially considering that genomic variation other than SNPs was not included in our analysis.

Part of the study data was provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161,R01AG15819,R01AG17917,U01AG46152. The other part of the data was provided by the Alzheimer's Disease Neuroimaging Initiative(ADNI).

Email: m.machnicka@mimuw.edu.pl

NGS data analysis using Hidden Markov Models

Ania Macioszek[1] and Bartek Wilczyński[1]

[1]Institute of Informatics, University of Warsaw

Experiments based on next generation sequencing are widely used in many different applications. They can be used to determine epigenomics landscape of specific types of cancer, find differences in expression between different tissues, characterize regulatory networks and others. In general, many of them - e.g. ChIP-seq, ATAC-seq, DNase-seq - produce a signal over the whole genome. Usually the following analysis include identifying some regions of interest basing on the value of the signal; e.g. in single ChIP-seq experiment, regions with enriched signal are cosidered to be sites of protein binding.We developed a tool to analyse such signals. The tool uses Hidden Markov Model with Gaussian emissions. It can be used to analyse one signal from a single experiment or many signals at once; in particular, one can use it on signals coming from different conditions (e.g. different tissues) to identify regions that differ between the two conditions.

Email: a.macioszek@mimuw.edu.pl

Tests and applications of Bioshell suite v.3.0

Joanna Magdalena Macnar[1,2], Natalia Anna Szulc[3,4], Aleksandra Elżbieta Badaczewska-Dawid[1], and Dominik Gront[1]

[1]Faculty of Chemistry, University of Warsaw
[2]College of Inter Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw
[3]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw
[4]International Institute of Molecular and Cell Biology in Warsaw

BioShell is a software toolkit for structural bioinformatics that has been developed for more that ten years. Its newest version, released on Apache 2.0 license and implemented in C++0x as well as in python, 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. Although some of the Bioshell functions, such as the contact map or the Ramachandran Plot, are also available in other Python libraries, such as BioPython, the tests performed so far indicate at least a 100-fold acceleration of calculations for large structures relative to Biopython.

Email: joanna.macnar@student.uw.edu.pl

FleXgeo: a machine learning-friendly representation of the universe of protein structures

Antonio Marinho da Silva Neto[1] and Stanisław Dunin-Horkawicz[1]

[1]Centrum Nowych Technologii, Uniwersytet Warszawski

A proper mathematical representation of protein structure space is essential for understanding its complexity and evolutionary processes underlying its emergence. Commonly used representations have significant limitations, for instance, 1) atomic coordinates relies on solving the structural superposition problem; 2) *phi-psi* angles are non-independent periodic variables, which is problematic for clustering, and 3) collective variables are not scalable. Considering the above, we explored the usage of a protein backbone representation based on differential geometry and knot-theory descriptors, implemented in FleXgeo package [1]. In FleXgeo, backbones are represented as regular 3D curves by cubic spline interpolation and then curvature and torsion values are computed for each residue. Moreover, the writhing number is computed to quantify the degree of backbone crossing for a given backbone region. The FleXgeo representation (i.e. torsion, curvature, and writhing number values) was calculated for >19k structures obtained from Protein Data Bank. Analysis of this dataset revealed highly-populated regions corresponding to known secondary structure elements (SSE) such as *alpha*, *pi*, and 310 helices, β-strands and PPII. However, we found many instances of SEE that are localized outside these regions, indicating that the structural space of SSE is not discrete but rather adopts a form of a continuum. We investigated the continuous nature of the SSE space in the context of PiPred, a machine learning-based tool for the prediction of *pi*-helices [2].

1. Silva Neto, A. M. et al. Proteins. ; 87: 302– 312 (2019) 2. Ludwiczak, J. and Winski, A. et al. Sci. Rep. 1–9 (2019)

Email: a.marinho@cent.uw.edu.pl

Allosteric signal transduction in neuronal protein reelin

Karolina Mikulska-Ruminska[1] and Wieslaw Nowak[1]

[1]Institute of Physics, Nicoulaus Copernicus University

Reelin is a large extracellular matrix glycoprotein which governs cell migration and positioning in the developing brain through activation of multiple intracellular signaling events by interactions with two lipoprotein receptors: ApoER2 and VLDLR, and with intracellular adaptor protein Dab1 [1].

In order to characterize crucial elements of reelin which may play an important role in singal transduction, perturbation scanning response (PRS) analysis of molecular dynamics (MD) simulations results was performed. PRS method has been used succesfully for other systems to predict the signal transduction pathways and potential receivers of allosteric signals [2,3]. We show that this approach helps to elucidate residues critical for reelin signal transduction through ApoER2, VLDLR and its other parthers.

Acknowledgements: Supported by Polish National Science Centre grant 2012/05/N/ST3/03178 and 2016/23/B/ST4/01770. We are thankful for the computer time allocated by the Interdisciplinary Center for Modern Technologies, NCU.

[1] E. Khialeeva et al., Develop. Dynamics 246, 4 (2017).

[2] C. Atilgan et al., Biophys. J. 99, 3 (2010).

[3] K. Mikulska-Ruminska et al., J. Chem. Inform. Model. 59 (2019).

[4] C. Quattrocchi et al., J. Biol. Chem. 277, 1 (2002).

Email: karolamik@fizyka.umk.pl

The application of Genetic Algorithm to predict the solvation energy parameters in extended coarse-grained method in implicit solvent environment

Przemysław Miszta[1], Szymon Niewieczerzał[1], Paweł Pasznik[1], Krzysztof Młynarczyk[1], and Sławomir Filipek[1]

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

Most of the biological processes occur on the timescales much too long to be investigated using all-atom simulations. Such processes like dynamics of large protein complexes or protein self-assembly are computationally very demanding and in order to study these phenomena one needs to introduce simplifications leading to the reduction of the degrees of freedom of the investigated system. This purpose can be achieved by simplifying the molecular representation of the system by replacing groups of atoms with effective grains, or by removing molecules to represent solvent implicitly.

We propose a method based on both mentioned approaches. The first of them is the MARTINI coarse-grained (CG) force field [1]. In this model one grain represents on average four heavy atoms, which originally, apart from protein structures, also applies to the solvent and lipids. The second approach is based on the Gaussian solvent-exclusion model. Originally this method provides the effective energy function called EEF1 for proteins in solution [2], and its extension to heterogenous membrane-aqueous media (IMM1) [3] is applicable to transmembrane proteins or proteins partly embedded in the lipid bilayer. Usage of CG representations in implicit environments allows to increase time of molecular dynamics simulations of membrane protein system by at least one order of magnitude or to increase number of proteins in the system by one order of magnitude, depending on size of protein. The method may be used for study of formation of protein oligomers and their dynamics in cell membranes.

The set of solvation energy parameters defining behavior of grains in the implicit aqueous-membrane media was determined using Genetic Algorithm (GA) in Python language. In this case the chromosome is an array of integer indexes representing floating point solvation parameters in a given range consisted of about 100 values. Four different solvation parameters for 50 CG bids grouped for similar parameters but calculated independently for water and membrane environments formed the chromosome with the number of parameters about 140 and the search space consisted of about 10^{200} variants. Employing GA we have obtained the best solvation parameters to mimic the implicit solvent in the force field that provide excellent stability of transmembrane proteins within 5Å of RMSD for 20 microseconds of coarse-grained molecular dynamics simulations.

[1] Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P. and Marrink, S.J., 2008. The MARTINI coarse-grained force field: extension to proteins. Journal of chemical theory and computation, 4(5), pp.819-834.

[2] Lazaridis, T. and Karplus, M., 1999. Effective energy function for proteins in solution. Proteins: Structure, Function, and Bioinformatics, 35(2), pp.133-152.

[3] Lazaridis, T., 2003. Effective energy function for proteins in lipid membranes. Proteins: Structure, Function, and Bioinformatics, 52(2), pp.176-192.

Email: pmiszta@chem.uw.edu.pl

New perspective in quadruplex classification

Joanna Miskiewicz[1], Mariusz Popenda[2], Joanna Sarzynska[2], Tomasz Zok[1,3], and Marta Szachniuk[1,2]

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

In recent years, quadruplexes became one of the most explored structural motifs in biological and bioinformatics fields. The need to understand the functions and capabilities of quadruplexes resulted in intensified study of their structure. It was performed mainly on the sequence and the tertiary structure level. In last decade, research group of Webba da Silva established quadruplex 3D structure classification based on glycosidic bond angles between the component nucleotides. They also defined quadruplex categories following the topologies of loops between consecutive G-tracts. This inspired us to create novel classification of quadruplexes based on their secondary structure.

Hereby, we present the results of our recent research on quadruplexes. We focused on the secondary structure of quadruplex motifs. Our study of graphical diagrams of tetrads created by RNApdbee resulted in defining their new classification named ONZ. ONZ allows to assign tetrads to three different categories based on the sequence order of each N-tract (from 5' to 3' end) and connections between N-tracts. We searched all the PDB-deposited nucleic acid structures to find 294 candidates that included at least one quadruplex. With the newly developed automated method, we categorized all tetrads and quadruplexes with reference to ONZ classes. We also proposed to extend the standard dot-bracket notation by including a 2-line format, to represent quadruplexes. We believe, that our classification will bring a new perspective in quadruplex analysis and it will contribute to a better understanding of these structural motifs.

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

Email: joanna.miskiewicz@cs.put.poznan.pl

A Comparative Study of Neuroactive Ligands Docking to Muscarinic M1 Type GPCR.

Beata Niklas[1], Karolina Mikulska-Rumińska[1], Bruno Lapied[2], and Wiesław Nowak[1]

[1]Nicolaus Copernicus University in Torun, Poland
[2]Univesity of Angers, France

Malaria is disease spread by mosquitoes that every year affects millions of people resulting in up to 730 thousand deaths. The most commonly used mosquito repellents (i.e. DEET) were found to be neurotoxic for humans, especially for children. There is a high need for new generation of repellents. Possible strategy is exploiting synergy effects of agents acting on G-protein coupled receptors (GPCRs). These membrane proteins participate in signal transduction and are targets of >30% of all modern drugs. By using molecular dynamics (MD) and docking tools we investigated possible docking places of a series of 5 ligands in two carefully selected structures of human M1 GPCR: (A) based on a homology modeling (B) determined by the X-ray crystallography. In this work we search for conformational changes induced by a ligand binding to the orthosteric and the allosteric sites of the M1 receptor. We compare Autodock VINA poses predicted for A and B structures and present the MD data (50 ns timescale) showing how docking of our neuroactive ligands affects fluctuations of each structure. The gather biostructural data may be used to feed some artificial intelligence or machine learning systems which in turn may help to elucidate mechanistic aspects of GPCR signaling and may lead to the development of better mosquito repellents.

Acknowledgements: support from the project POWR.03.05.00-00-Z302/17 Universitas Copernicana Thoruniensis in Futuro-IDS "Academia Copernicana" Centre of Informatics - Tricity Academic Supercomputer & network (CI TASK)

Email: beata.niklas.x@gmail.com

The Polish Bowel Sound Group: A Plan for Action

Jan K. Nowak[1], Przemysław Szelejewski 1[2], Stefan Stróżecki[3], Marcin Dziekiewicz[4], and Jarosław Walkowiak[1]

[1]Department of Pediatric Gastroenterology and Metabolic Diseases, Poznan University of Medical Sciences, Poznan, Poland (PUMS) 1
[2]EPS, Poznan, Poland 2
[3]Institute of Informatics and Mechatronics, University of Economy, Bydgoszcz, Poland 3
[4]Department of Pediatric Gastroenterology and Nutrition, Medical University of Warsaw, Warsaw, Poland 4

Background: Bowel sounds (BS) are ubiquitous yet under-researched. We aim to aggregate a large number of recordings from a heterogenous cohort in order to facilitate further BS research. Methods: The construction of dedicated abdominal contact microphones will soon be completed. Portable 24-bit linear recorders will enable the capture of signals within a high dynamic range. The Polish Bowel Sound Group is being formed, with the intention of including more than six clinical centers and experts in related fields, such as bioinformatics, data analysis, and acoustics. Healthy children and adults as well as patients with different gastrointestinal diseases will be enrolled. The subjects will be instructed to attach the microphone to the right lower quadrant of the abdomen immediately prior night sleep and to continue the recording until the first meal in the morning is finished. Clinical data will be gathered through a self-administered questionnaire. The study was approved by the Bioethical Committee at PUMS (no. 683/18). Expected results: The Polish Bowel Sound Group intends to aggregate a large number of recordings accompanied by essential clinical data. Efficacious BS identification will constitute a major challenge. Fundamental hypotheses regarding BS will be tested. Ideally, the further growth of the database should allow the Group to reach the momentum needed to stimulate BS research and allow new practical applications. The Group intends to hasten progress in the field of BS by the sharing of equipment, knowledge, and data. Conclusions: Because of technical maturity, BS research is at a turning point worldwide. The Polish Bowel Sound Group intends to gather and share the largest BS collection to date. Interested researchers are welcome! (PUMS grant 21/MN/2017)

Email: jan.nowak@ump.edu.pl

Challenge of classification of patients with Alzheimer's disease based on DNA polymorphisms

Marlena Osipowicz[1], Magdalena Machnicka[1], and Bartek Wilczyński[1]

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

Solving the issue of genotype-based classification of patients suffering from Alzheimer's disease (AD) may provide us valuable information about DNA regions functionally relevant for this disorder. Several machine learning approaches have been already applied to this problem and reported classification accuracies ranging from 58The classification process consists of three parts: feature selection, classification and evaluation. In this project the Boruta feature selection algorithm was used together with the random forest classifier and AUROC metric. Three datasets were analyzed: WGS and GWAS data collected by the Alzheimer's Disease Neuroimaging Initiative (ADNI) and WGS data from the Religious Orders Study and the Rush Memory and Aging Project (ROS/MAP). In total, genetic data of 1519 patients was used (765 cases and 754 controls). The main analysis was performed in two different ways: division into training and testing sets was conducted after selection of the most important SNPs (resulted AUROC 0.98) or before (resulted AUROC 0.6). For the reasons mentioned above the first score is apparently a result of overfitting. Further modifications of the model such as merging datasets resulted in slight increase in the performance. As the next step we plan to investigate relevant features selected by the model with respect to their potential functionality in AD.

Email: marlena.osipowicz@gmail.com

Analysis of EGFRvIII ectodomain conformational transition with Targeted Molecular Dynamics

Marcin Pacholczyk[1,3] and Piotr Rieske[2,3]

[1]Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
[2]Department of Tumor Biology, Medical University of Lodz, Lodz, Poland
[3]Research and Development Unit, Celther Polska Ltd., Lodz, Poland

EGFR extracellular domain spanning over 600 amino acids has two functionally important conformations (or states) – inactive, autoinhibited or tethered (T) monomer and ligand bound, active or untethered (UT) form which promotes EGFR dimerization. Deletion of exons 2-7 of wild type EGFR causes presence of unpaired cysteine in EGFRvIII monomer. Previous works suggested that Cys16 is unpaired can directly participate in covalent homodimerization of EGFRvIII [1], whereas other cysteines form disulfide bonds within monomers. Despite mutation, hinge region responsible for conformational transition between T and UT states remains intact in EGFRvIII mutant. Crystallographic model of EGFRvIII covalent homodimer is not available. Homology model based on EGFR homodimer (PDB ID: 3NJP) does not show conformation suitable for covalent homodimerization trough C16. The main goal of this work was to simulate conformational transition between UT and T states to capture conformation in which EGFRvIII most likely homodimerize covalently. Timescales of large conformational transitions are usually unavailable to classical Molecular Dynamics (MD). Additional force applied during TMD allows the system to cross energy barriers and the conformational transition can be observed in tens of nanoseconds of simulation. Using TMD simulation (30 ns) trajectory of conformational transition between UT and T states was registered. EGFRvIII covalent homodimerization is most likely a dynamic process and requires some conformational adaptation of the ectodomain in the hinge region.

1. Stec W. et al. Cyclic trans-phosphorylation in a homodimer as the predominant mechanism of EGFRvIII action and regulation. Oncotarget. 2018; 9:8560-8572

Email: marcin.pacholczyk@polsl.pl

Drug-target genes and drugs identification for more effective diagnosis and treatment of the squamous-cell carcinoma and adenocarcinoma esophageal cancers

Aneta Polewko Klim[1], Sibo Zhu[2], Xingdong Chen[3], and Witold R. Rudnicki[1]

[1]Faculty of Mathematics and Informatics, University in Bialystok, Poland
[2]Department of Epidemiology, School of Public Health, Fudan University , China
[3]Fudan-Taizhou Institute of Health Sciences, China

Despite that two main histological subtypes of esophageal cancer (SCA) i.e. esophageal squamous cell carcinoma (SCC) and esophageal adenocarcinoma (AC) cancer are increasingly considered as separate diseases, the recommendations of chemotherapy treatment and biological target medical therapies are similar for both types. The identification of the drug-genes potentially relevant for a different response to treatment in SCC and AC patients is crucial for a more effective treatment of these two groups of patients. In this paper, we present the investigation of the genes encoding membrane proteins related to anti-cancer drugs, belonging to group of the most informative biomarkers which can be applied to distinguish SCC and AC by using advanced machine learning algorithms. Predictive models were built using Random Forest classification algorithm using RNA-seq data from The Cancer Genome Atlas data portal project integrated with data deposited in The National Omics Data Encyclopedia of China collected at the Fudan-Taizhou Institute of Health Sciences. The ensemble of five different filter feature selection methods (Welch t-test, the one-dimensional and two-dimensional feature selection filter based on information theory (MDFS), the fast correlation-based filter (FCBF), the ReliefF algorithm and the minimum redundancy and maximum relevance (mRmR) were used for discovering the most important genes. Integration of the classification results and pharmaceutical database information, allowed us to identify four drug-related genes, namely ERBB3, ATP7B, ABCC3 and GALNT14, as potentially important for a more effective immunotherapy and chemotherapy of SCA patients. The pathways analysis was performed to investigate functional role of the most infor

Email: anetapol@uwb.edu.pl

Mobile monitoring system for heart disease

Sandra Śmigiel[1]

[1]UTP University of Science and Technology

Electrocardiograms (ECGs) can provide useful information about the physical condition of a patient. This paper proposes the design of a household device that will allow detecting T-wave abnormalities. This device consists of two sub-devices and can indicate the risk of a heart attack. The first sub-device collects electrocardiographic data and communicates with the second subdevice wirelessly. The second sub-device carries out pre-filtered signal analysis and determines if there is a threat. A standard mobile phone is used for the analysis and decision-making. An ECG analysis algorithm based on matched filtering in the time domain is also proposed. Filter pulse characteristics correspond to the distortion in the T-wave; this correspondence can be used to obtain information about its deficiencies after analysing the effect of the main lobe to side lobes factor in the result of the convolution. Two additional algorithms were used to support the analysis and to increase its effectiveness. The first algorithm is responsible for matching the impulse characteristics of the filter to the current heart rate of the tested patient. The second algorithm standardizes the amplitude values in the analysed signals. In the study, the PhysioBank database was used for input data. The developed algorithms were tested on 50 selected ECG signals of patients with previous myocardial infarction and 10 signals which came from healthy patients. The decision algorithm arrived at correct decisions with an efficiency of more than 90%. The proposed solution can help improve current statistics and increase the patient's chances of survival.

Email: sandra.smigiel@utp.edu.pl

The Machine Learning approach to recognize the laterality in real and virtual test tasks

Beata Sokołowska[1], Ewa Sokołowska[2], Teresa Sadura-Sieklucka[3], Magdalena Łachwa-From[1], Monika Górecka[1], Dagmara Kabzińska[1], Weronika Rzepnikowska[1], Katarzyna Binięda[1], Artur Kiepura[1], Anna Sobańska[4], and Małgorzata Dylewska[5]

[1]Mossakowski Medical Research Centre Polish Academy of Sciences, Warsaw, Poland
[2]John Paul II Catholic University of Lublin, Lublin, Poland
[3]National Geriatrics, Rheumatology and Rehabilitation Institute, Warsaw, Poland
[4]Institute of Psychiatry and Neurology, Warsaw, Poland
[5]Institute of Biochemistry and Biophysics Polish Academy of Sciences, Warsaw, Poland

The study presents the results of healthy volunteers in tests using (1) Piórkowski's real device and (2) virtual cognitive-motor tasks. Each task was realized twice with the right and left hand, independently. In Piórkowski's tests, three different exposures of light stimuli were used with slow, fast and very fast speeds. The virtual tasks with the recognition of Colored Balls were also carried out on three levels of difficulty/ duration: easy, difficult and very difficult. The aim of the research was to estimate and to recognize the laterality of subjects using the Machine Learning algorithms based on pattern recognition methods. Differentiation of the performance of the tasks by the dominant and non-dominant hand was made for the real measuring instrument (lighting up lamps) and virtual objects (selection of appropriate balls). Better results of the dominant hand were observed in both the real (RT) and the virtual tests (VT). In addition, the non-dominant hand showed more signs of fatigue for RT than for VT. The subjects were more engaged in achieving better results in VT compared to RT. In summary, innovative virtual exercises are very attractive, interactive and also effective in assessing the level of effective in assessing the level of hand tasks. Authors wish to acknowledge researches and students who participated in this project, in particular to Prof Lesyng B, Dr Krzyśko K, Charzewski Ł, Firlik G, Komar M and Błażejewska A from UW, also to Dr Czerwosz L, Dr Kolinski M, Dr Rakowski F from MMRC PAS, and to Siergiej Alicja from University of Perugia. This study was carried out using the computational infrastructure of the Biocentrum-Ochota project (POIG.02.03.00-00-0030/09) and applying NEUROFORMA's virtual software of Titanis' technologies.

Email: beta.sokolowska@imdik.pan.pl

SamCC-Turbo: high-throughput and precise measurement of coiled-coil protein domains

Krzysztof Szczepaniak[1] and Stanislaw Dunin-Horkawicz[1]

[1]Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland

Coiled coils are widespread protein domains involved in diverse cellular processes ranging from providing the structural rigidity to transduction of conformational changes [1]. Coiled coils comprise two or more *alpha*-helices that are wound around each other to form a very regular braid-like supercoiled bundle. The regularity of coiled coils makes it possible to represent them using parametric equations and thus to derive basic parameters such as the orientation of helices, degree of supercoiling of the bundle and others. A coiled-coil bundle can be "sliced" into layers where each layer is formed by one residue from each helix. Layers are roughly perpendicular to the bundle axis and parallel to each other. Existing software can measure coiled-coil parameters for individual layers thus providing us with a detailed structural description of a given structure [2]. However, this program requires manual annotation of layers, hampering any high-throughput analysis. Here we present SamCC-Turbo software that aims at overcoming these difficulties. SamCC-Turbo takes as an input a protein structure, uses Socket [3] for the rough annotation of coiled-coil regions, calculates optimal layer setting, and finally, performs measurement of bundle parameters at per-layer resolution. SamCC-Turbo will facilitate large scale analysis of coiled coils which are an important model for studies of the relationship between protein sequence and structure [4, 5].

[1] Lupas AN, Gruber M. Adv Protein Chem. 2005;70:37-78. [2] Dunin-Horkawicz S, Lupas AN. J Struct Biol. 2010 May;170(2):226-35. [3] Walshaw J, Woolfson DN. J Mol Biol. 2001 Apr 13;307(5):1427-50. [4] Grigoryan G, Degrado WF. J Mol Biol. 2011 Jan 28;405(4):1079-100. [5] Szczepaniak K et al. J Struct Biol. 2018 Oct;204(1):117-124.

Email: k.szczepaniak@cent.uw.edu.pl

HiCEnterprise: Identifying long range chromosomal contacts in HiC data

Irina Tuszynska[1], Hanna Kranas[1], and Bartek Wilczynski[1]

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

HiCEnterprise is a software tool for identification of long-range chromatin contacts based on the Hi-C experiments. It implements four different statistical tests for identification of significant contacts at different scales as well as necessary functions for input, output and visualization of chromosome contact matrices. HiCEnterprise allow identifying chromosomal contacts on the level of enhancer-promoter interactions using Gumbell distribution (Won et al., 2016) as well as domain-to-domain interactions by hypergeometric (Niskanen et al., 2017), Poisson and negative-binomial distributions.

Email: irina@mimuw.edu.pl

Metadynamics as a tool for a local perturbation in diabetes related large protein complexes

Katarzyna Walczewska-Szewc[1,2] and Wiesław Nowak[1,2]

[1]Department of Physics, Astronomy and Applied Informatics, Nicolaus Copernicus University
[2]Centre for Modern Interdisciplinary Technologies, Nicolaus Copernicus University

Metadynamics is the approach of the sampling improvement based on the local modification of the potential energy landscape. Motion of the system along the configurational space of relevant collective variables (CV) can be sped up by adding an artificial Gaussian-shaped bias potential to the energy in each step, to discourage the system from visiting previously visited states. The key to metadynamic simulations is the choice of a proper set of collective variables. These should be sufficient to describe the process we want to study, but may be difficult to sample due to the existence of potential energy barriers or just too slow dynamics along the CV during a simulation.

Here we use the metadynamics approach to model computationally the system of SUR1 (the regulatory unit of ATP-sensitive potassium channel, KATP) with the externally controlled sulfonylurea derivative docked to its pocket. Conventional sulfonylurea drugs are often used to control the blood glucose level in type 2 diabetes (T2D) patients. The role of such drugs is to inhibit the action of KATP which triggers the signalling pathway leading to the insulin release. Metadynamics is used to drive the embedded drug between two states showing the structural rearrangements of the whole system induced by the external factor. Such rearrangement can potentially lead to the on/off signaling of KATP system.

Email: kszewc@fizyka.umk.pl

Polymer molecule under thermodynamic forces

Piotr Weber[1]

[1]Department of Atomic, Molecular and Optical Physics, Faculty of Applied Physics and Mathematics, Gdańsk University of Technology

Dynamics of polymer molecules can be given at various levels of detail. The most detailed descripion we can obtain from molecular dynamics, which is based on classical mechanical approach. Apart of this, there are methods that decribe polymeric system in mesoscopic level, where coarse-grained models are used. In this approach microscopic degrees of freedom are eliminated and a polymer molecule is represented by a simplified structure - a chain.

Using statistical mechanical approach, the set of elementary events consist of different configurations of a macromolecule. Statistics of molecule configurations depend on their intrinsic property, the interactions between their different parts and on the interactions with the environment. This environment can be the source of thermodynamic forces, that affect the polymer molecule. Variations in conformation of the macromolecules can be described by the mesoscopic nonequilibrium thermodynamics to analyze the irreversible processes taking place in the conformational space of the macromolecules. This theory can also be used in the case to derive the kinetic equations of polymer dynamics.

The poster presents a consideration about polymer under thermodynamic forces and the idea that it can be consider as a fractional free enthalpy transducers.

Email: piotr.weber@pg.edu.pl

Assessing similarity in the set of RNA 3D structures

Jakub Wiedemann[1,2], Tomasz Zok[1,3], Maciej Milostan[1,2], and Marta Szachniuk[1,2]

[1]Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology
[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Z. Noskowskiego 12/14, 61 704 Poznan, Poland
[3]Poznan Supercomputing and Networking Center, Jana Pawła II 10, 61 139 Poznan, Poland

Identification of common features and differences in biomolecule structures is one of the most important tasks in the field of bioinformatics. With a growing number of available structures predicted both in silico and experimentally, there is still a need to have the ability to compare them. Most of the current approaches are not suitable to compare sets of structures and find common elements within them. Usually, the limitation boundary is reached at a pairwise comparison of the structures. Hereby, we propose a new approach that allows finding the longest continuous segments in 3D RNA structures in the set using torsional angle representation (MultiLCS-TA). An advantage of the method is its independence on structural alignment which decreases the computational complexity of the whole process.

Acknowledgments

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

Email: jakub.wiedemann@cs.put.poznan.pl

PiPred – a deep-learning method for prediction of *pi*-helices in protein sequences

Jan Ludwiczak[1,2], Aleksander Wiński[1], Antonio Marinho da Silva Neto[1], Krzysztof Szczepaniak[1], Vikram Alva[3], and Stanislaw Dunin-Horkawicz[1]

[1]Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097, Warsaw, Poland
[2]Laboratory of Bioinformatics, Nencki Institute of Experimental Biology, Pasteura 3, 02-093, Warsaw, Poland
[3]Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, Max-Planck-Ring 5, 72076, Tubingen, Germany

pi-helices are short and unstable protein secondary structure motifs. They are present in 15% of all known protein structures, often in functionally important regions such as ligand-binding sites. Thus, the correct prediction of *pi*-helices is essential for the detection of potential functional sites and can aid design tasks aiming at the creation of ligand-binding pockets. Due to the similarity of *pi*-helices to much more abundant *alpha*-helices, it is a challenging task to predict them based on the sequence data and there are no methods devoted to this problem. We present a deep learning neural network trained with sequences where *pi*-helical residues were assigned based on high-quality X-ray structures [1]. The model achieved 48% precision and 46% sensitivity in per-residue prediction on a test set. Moreover, in the benchmark on commonly used datasets like CB6133, CB513, CASP10, and CASP11 the model outperforms other state-of-the-art methods used for secondary structure prediction. The developed tool called PiPred is freely accessible at https://toolkit.tuebingen.mpg.de/#/tools/quick2d. A standalone version is available for download at https://github.com/labstructbioinf/PiPred.

[1] Jan Ludwiczak, Aleksander Winski, Antonio Marinho da Silva Neto, Krzysztof Szczepaniak, Vikram Alva & Stanislaw Dunin-Horkawicz PiPred – a deep-learning method for prediction of *pi*-helices in protein sequences. Scientific Reports, 2019, May 3

Email: a.winski@cent.uw.edu.pl

Bioinformatic characteristic of novel protein superfamily DM9/MFP2

Tuguldur Enkhbaatar[1], Joanna Ziemska-Legięcka[2], and Marcin Grynberg[3]

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

The DM9 domain was first described in 2001 by Ponting and colleagues as ubiquitous was first noticed in the model organism Drosophila melanogaster. It consists of a repeated motif, which always appears in even numbers. This 60–75 residues long motif is commonly found in Arthropoda, especially in insects. Recently, a link between a trematode Mfp2 protein and the DM9 proteins was revealed. Mfp2 is involved in cytoplasmic motility, whereas the DM9-containing protein CGL1 from the oyster is known to be a lectin that binds mannose. In addition, recent studies indicate that the PRS1 protein, which contains the DM9 domain, contributes to the interaction of Plasmodium with the mosquito epithelium. This interaction does not depend on the specific co-evolutionary history of these two organisms.

In our research we use computational biology methods to undertake a large DM9 and Mfp2 superfamily search in order to closely define this ideal tandem repeat domain taxonomically, phylogenetically and structurally. We prove that DM9 is present in many taxonomic lineages and show unexpected functional correlations of this domain family.

Email: jz344988@students.mimuw.edu.pl

Workshops

Enhanced Sampling Method for Ligand Unbinding

Jakub Rydzewski[1]

[1]Institute of Physics, Nicolaus Copernicus University, Grudziadzka 5, 87-100, Torun, Poland

Recent developments in enhanced sampling methods showed that it is possible to find ligand unbinding pathways with spatial and temporal resolution inaccessible to experiments. Such techniques should provide an atomistic definition of possibly many reaction pathways, otherwise it may lead either to overestimating energy barriers, or inability to sample hidden energy barriers that are not captured by simple reaction pathway estimates. We provide an official Plumed 2 module that implements a method which is able to sample the reaction pathways of the ligand-protein dissociation process. The method is based on enhanced sampling and non-convex optimization methods. The module, called MAZE, requires only a crystallographic structure to start a simulation, and does not depend on many ad hoc parameters. To present its applicability and flexibility, we provide several examples of ligand unbinding pathways along transient protein tunnels reconstructed in a model ligand-protein system.

Email: `jr@fizyka.umk.pl`

Building robust machine learning models

Aneta Polewko-Klim[1], Krzysztof Mnich[2], Wojciech Lesiński[1], Radosław Piliszek[2], Bogumił Sapiński[3], and Witold Rudnicki 1 2[3]

[1]Institute of Informatics, University of Białystok,
[2]Computational Centre, University of Białystok,
[3]Interdisciplinary Centre for Mathematical and Computational Modelling,

Background: Modern experimental techniques deliver data sets containing profiles of tens of thousands of potential molecular and genetic markers that can be used to improve medical diagnostics. We propose methodology arise due to limited sample size and feature selection. It is based on comprehensive cross-validation protocol, that includes feature selection within cross-validation loop and classification using machine learning. This methodology allows for estimation of biasaes inherent in the machine learning. Protocol was utilised for building several models based on sets of variables of varying sizes that were selected using three different feature selection methods.

Results: The significant biases due to feature selection procedure and split of the sample between training and validation sets were observed. The size of the bias depends on the size of experimental sample. Good correlation between performance of the models in the internal and external cross-validation was observed, confirming the robustness of the proposed protocol and results.

Conclusions: We have developed a protocol for building predictive machine learning models. The protocol can provide robust estimates of the model performance on unseen data. It is particularly well-suited for small data sets. We have applied this protocol to develop prognostic models for neuroblastoma, using data on copy number variation and gene expression.

Email: W.Rudnicki@icm.edu.pl

The coarse-grained UNRES force field description and usage of the UNRES server in modeling of protein structure

Agnieszka S. Karczyńska[1], Agnieszka G. Lipska[1], Emilia A. Lubecka[2], Adam K. Sieradzan[1], Cezary Czaplewski[1], and Adam Liwo[1]

[1]Faculty of Chemistry, University of Gdansk, Wita Stwosza63, 80-308 Gdańsk, Poland
[2]Institute of Informatics, Faculty of Mathematics, Physics, and Informatics, Wita Stwosza 57, 80-308 Gdańsk, Poland

The UNRES package (http://www.unres.pl) is an implementation of the physcis-based UNRES coarse-grained model of polypeptide chains [1], which can be used in molecular simulations, including those aimed at protein-structure prediction. The UNRES model has only two interaction sites per amino-acid residue, namely a united side chain and a united peptide group, the solvent present implicitly in the effective potentials. This reduction of polypeptide chain representation enables us to run relatively long simulations for big protein systems in reasonable time, without ancillary information from structural databases. Nevertheless, the implementation includes the possibility of using restraints derived from structural databases or experimental data. Recently [2], we developed a web served based on the UNRES package, freely available at http://unres-server.chem.ug.edu.pl. The following tasks can be run using the server: (i) local energy minimization, (ii) canonical molecular dynamics simulations, (iii) replica exchange and multiplexed replica exchange, which are aimed at the simulation of conformational ensembles and protein structure prediction. The user-supplied input includes protein sequence and, optionally, restraints from secondary-structure prediction or small x-ray scattering data [3], and simulation type and parameters which are selected or typed in. Oligomeric proteins, as well as those containing D-amino-acid residues and disulfide links can be treated. The output is displayed graphically (minimized structures, trajectories, final models, analysis of trajectory/ensembles, residue-fluctuation profiles) and all output files can be downloaded by the user in the plain text version.

Email: `lypstyk@gmail.com`