X Symposium of the Polish Bioinformatics Society

27–29.09.2017, Uniejów, Poland

http://ptbi2017.cs.put.poznan.pl

BOOK OF ABSTRACTS

Dear Colleagues,

I have the great pleasure to welcome you all at the tenth Symposium of the Polish Bioinformatics Society (PBS). For the first few meetings we used the less formal name of Convention, however, the increasing quality of the meetings prompted us to adopt the current name, beginning with the fifth meeting.

The first Convention took place in 2008 in Jadwisin near Warsaw and was organized by a team from the Institute of Informatics, University of Warsaw. The meeting was attended by almost 60 participants. The convention was a continuation of informal PhD workshops previously organized by the teams of Jerzy Tiuryn (University of Warsaw) and Jacek Błażewicz (Poznan University of Technology). We have inherited not only the legacy but also a formula of the meeting. PhD students and MSc students as well as freshly minted Doctors presented their research orally, with ample time for discussion. All subsequent events were built on that tradition.

The Symposium is a place where young bioinformatics practitioners present their research on a forum broader than their own laboratory or institute community. Symposia have a great significance for the integration of the bioinformatics community in Poland. During our short

history, they were organized in (or in the vicinity of) Warsaw (Jadwisin 2008), Poznan (Będlewo 2009), Gliwice (Ustroń 2010), Kraków (2011), Gdańsk (2012) and Wrocław (2013). In 2014, the Symposium returned to Warsaw as an integral part of the first BIO convention organised by four scientific communities - biochemists, biophysicists, cell biologists and bioinformaticians. In 2015, we organized the Symposium in Lublin, together with the Polish Society of Medical Chemistry. In 2016 the symposium took place in Białystok. Current edition is again organized by the community from Poznan University of Technology (though it takes place near Łódź). In 2018 the Symposium will return to Wroclaw.

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

Our Symposia are changing with us. In 2008 the community was very young. Most participants in the Congresses were students, doctoral students and young doctors. But time does not stand still. From young doctors, young habilitated doctors were born, previous

students are now doctors, who are not so young anymore. Symposia reflect this chance. Therefore, we have introduced opportunity to present for more experienced colleagues.

The Symposium is also an opportunity to get acquainted with interesting achievements in our broadly understood scientific neighbourhood. To the event we prominent local scholars from related fields, as well our distinguished foreign collaborators.

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

On behalf of the PBS Board I would like to thank members of the Organizing and Program Committees for the impeccable organization of the Symposium. I hope that the jubilee tenth edition will be the best one yet. To the presenters I wish successful presentations and excellent posters. To all participants I wish excellent atmosphere and fruitful scientific discussions during the entire conference, the pool session included.

Witold Rudnicki
President of the Board
Polish Bioinformatics Society

PROGRAM COMMITTEE

MARTA KASPRZAK[1] – CHAIR

PIOTR FORMANOWICZ[1]– CHAIR

WITOLD RUDNICKI[2]

AGNIESZKA RYBARCZYK[1]

MARTA SZACHNIUK[1]

PAWEL WOJCIECHOWSKI[1]

ORGANISING COMMITTEE

MARTA SZACHNIUK[1] – CHAIR

MACIEJ ANTCZAK[1]

MARCIN BOROWSKI[1]

ALEKSANDRA GRUCA[3]

MARCIN RADOM[1]

KAROLINA ZWIEWKA[1]

1. POZNAN UNIVERSITY OF TECHNOLOGY, POLAND
2. UNIVERSITY OF BIAŁYSTOK, POLAND
3. SILESIAN UNIVERSITY OF TECHNOLOGY, POLAND

SPONSORSHIP

An organization of the X Symposium of the Polish Bioinformatics Society has been supported by:

Ministerstwo Nauki i Szkolnictwa Wyższego

Polish Ministry of Science and Higher Education
(under grant no. 842/P-DUN/2017)
http://www.nauka.gov.pl

Polish Bioinformatics Society
http://www.ptbi.org.pl

Institute of Computing Science,
Poznan University of Technology
http://www.cs.put.poznan.pl

EWG CBBM COMPUTATIONAL BIOLOGY BIOINFORMATICS AND MEDICINE

EURO working group on Computational Biology, Bioinformatics and Medicine
http://euro-cbbm.ku.edu.tr

CONTENTS

SCHEDULE OVERVIEW

WEDNESDAY, September 27

12:00 - 13:45	Registration and lunch
13:45 - 14:00	Conference opening
14:00 - 15:40	Session W1
15:40 - 16:00	Coffee break
16:00 - 17:40	Session W2
18:00 - 21:00	Poster session and buffet
19:00 - 20:00	PTBI Board meeting (Board members only)

THURSDAY, September 28

07:30 - 08:00	Jogging with the treasurer
09:00 - 10:40	Session T1
10:40 - 11:00	Conference photo and coffee
11:00 - 13:00	Social swimming
13:00 - 14:00	Lunch
14:00 - 15:40	Session T2
15:40 - 16:00	Coffee break
16:00 - 17:10	Session T3 (Awards session)
17:30 - 18:30	PTBI Convention
19:00 - 22:00	Conference dinner and awards

FRIDAY, September 29

09:00 - 10:40	Session F1
10:40 - 10:50	Closing remarks and farewell
10:50 - 11:30	Coffee and lunchbox
11:00 - 12:30	Session F2 (Workshops)

KEYNOTE SPEAKERS

JACEK BŁAŻEWICZ

Institute of Computing Science
Poznan University of Technology,
Poznan, Poland

Session F1: Friday, 29.09, 9:00

MACIEJ KOMOSIŃSKI

Institute of Computing Science
Poznan University of Technology,
Poznan, Poland

Session T2: Thursday, 28.09, 14:00

DIRK LABUDDE

Bioinformatics and Forensic Science
Investigation Laboratory
University of Applied Sciences,
Mittweida, Germany

Session W2: Wednesday, 27.09, 16:00

THOMAS VILLMANN

Computational Intelligence Group
University of Applied Sciences,
Mittweida, Germany

Session T1: Thursday, 28.09, 9:00

CONFERENCE VENUE

Location

The conference will be hosted in the 1600s Renaissance-style castle (Zamek Arcybiskupów Gnieźnieńskich) located at Turecka 12, 99-210 Uniejow (Lodz Voivodeship, Central Poland). GPS: N 51.97302778°, E 18.78880833°.

Transport

The most convenient way of getting to the venue is by own car, bike or canoe. However, using public transport (shuttle bus) is also possible.

Shuttle bus to Uniejów (PKS) leaves from Łódź Fabryczna Railway Station. One bus is available daily:

14:10 ŁÓDŹ – UNIEJÓW

Additionally, a special bus for conference participants is planned from Łódź Fabryczna Railway Station. Please, email the organizers to submit your interest in the conference bus.

Accommodation

All participants of the PTBI Symposium 2017 are kindly requested to make a hotel reservation (with payment) on their own. Rooms in the castle (**ZAMEK ARCYBISKUPÓW GNIEŹNIEŃSKICH**) and in **DOM PRACY TWÓRCZEJ** should be reserved with password "konferencja PTBI" before 31.07.2017.

UNIEJÓW ARCHBISHOP'S CASTLE

Uniejów Castle is one of the most interesting sight of historic, architectural and scenic value in the Łódź Voivodeship, Poland. It is famous for its long history and structure that picturesquely blends with the surrounding landscape: a river and its wetlands, a park - all together creating a historic and natural landscape complex.

The Castle was built in the years 1360-1365 replacing an old wooden fortification that was destroyed during the wars with the Order of Teutonic Knights in 1331. The fortified castle was erected on the initiative of Abp. Skotnicki, close associate to Casimir III the Great.

The building was greatly expanded and modernised between 1525-1534, when after a fire most of the castle's Gothic characteristics had gone. The stronghold had ended its militaristic significance in the seventeenth century, when in 1836 the castle became a residence of Estonian family, House of Toll. The Tolls had reconstructed the castle into a classical architectural style and remained the owners till 1918. During the interwar period the castle buildings served as a guesthouse. After WWII, it was used as a warehouse for fertilizers and crop. Only in 1957-1967 restoration and adaptation works were executed based on the design by H. and I. Ziętkiewicz.

SOCIAL PROGRAM

WEDNESDAY, September 27

Poster session and welcome buffet in the Herbowa Restaurant of Uniejow Castle at 18:00.

We are looking forward to seeing you in Castle of Uniejow! Our welcome reception on September 27th will take place in the atmospheric chambers and terrace of gothic Uniejow Castle – in the heart of Poland, only 50km from its geographical center.

That evening we would like to invite you to the local's chef show during welcome buffet that will accompany poster session. The cuisine of the castle Herbowa Restaurant combines modern nutritional trends with old Polish recipes. We hope that the chef's culinary skills, original atmosphere of the dining room, aroma of high quality products with natural herbs and spices, will make your evening unforgettable.

THURSDAY, September 28

Jogging, social swimming and anniversary dinner.

On Thursday morning, all jogging enthusiasts will meet at 7:30 in front of Uniejow Castle to enjoy running with the treasurer. Jogging will take place in the beautiful 19th century park surrounding the castle.

From 11:00 till 13:00, in the interval between Sessions we would like to invite you to participate in social swimming in the Uniejow Thermal Spa and Pool Complex located at the foot of the castle.

The Pool Complex features new indoor and outdoor pools as well as a spa centre. Outdoor pools have been connected to the indoor pool building and supplied with thermal brine water. Such a solution enables operation of the pools in any weather throughout the whole year. The Complex also boasts a brine pool with warm water and a separate outdoor pool divided into the main and children's sections. It is a place where everyone can find something interesting: whether hot or cold. A wide range of different saunas, a massage room, a hot brine pool and an ice pool, as well as a snow chamber will not only offer an unforgettable experience, but also leave your body relaxed and full of energy.

Day two of our conference will culminate in the Award Dinner and 10th Anniversary Ceremony. The dinner will have a medieval feast character and will be held in Herbowa Restaurant starting from 19:00. This evening highlight will be the PTBI Awards for best conference presentation and poster but, considering this unique opportunity which is the decade of PTBI activity, we have prepared also a special surprise.

SESSION OVERVIEW

WEDNESDAY, September 27

12:00 - 13:45	**Registration and lunch**
13:45 - 14:00	**Conference opening**
14:00 - 15:40	**Session W1** Chair: Małgorzata Kotulska
14:00 - 14:20	Paweł P. Woźniak, *Wroclaw University of Science and Technology, Poland* Forecasting Residue - Residue Contact Prediction Accuracy
14:20 - 14:40	Aleksandra I. Jarmolińska, *University of Warsaw, Poland* GapRepairer - Repair Protein Structures and Their Topology
14:40 - 15:00	Tymoteusz Oleniecki, *University of Warsaw, Poland* CABS-flex Standalone Application for Fast Simulations of Flexibility of Globular Proteins
15:00 - 15:20	Sebastian Bittrich, *Mittweida University of Applied Sciences, Germany* eQuant: a Web Service for Energy-Based Model Quality Assessment
15:20 - 15:40	Alexander Eisold, *Mittweida University of Applied Sciences, Germany* In Silico Method for Transformation of DNA/ RNA Aptamers into Peptide Nucleic Acid Aptamers
15:40 - 16:00	**Coffee break**

16:00 - 17:40	**Session W2** Chair: Aleksandra Gruca
16:00 - 16:40	**Keynote speaker:** Dirk Labudde, *Mittweida University of Applied Sciences, Germany* Watson meets Watson – How Today's Life Science Technologies Can Shape the Crime Sciences of Tomorrow
16:40 - 17:00	Radosław Piliszek, *University of Bialystok, Poland* MDFS- a Software Library for Multidimensional Feature Selection in Search for Relevant Genes
17:00 - 17:20	Małgorzata Marszałek-Zeńczak, *Institute of Bioorganic Chemistry PAS, Poland* Population-Scale Detection of Copy Number Variations in Arabidopsis Thaliana Genome
17:20 - 17:40	Tomasz Żok, *Poznan University of Technology, Poland* New Approaches to Determine RNA Pseudoknot Order
18:00 - 21:00	**Poster session and buffet**
19:00 - 20:00	PTBI Board meeting (Board members only)

THURSDAY, September 28

07:30 - 08:00	**Jogging with the treasurer**
09:00 - 10:40	**Session T1** Chair: Wiesław Nowak
09:00 - 09:40	**Keynote speaker:** Thomas Villmann, *Mittweida University of Applied Sciences, Germany* Classification and Pattern Recognition in Bioinformatics by Prototype-Based Machine Learning Models

09:40 - 10:00	Florian Kaiser, *Mittweida University of Applied Sciences, Germany* Mining Functionally Conserved Building Blocks in Biological Macromolecules
10:00 - 10:20	Aleksandra Suwalska, *Silesian University of Technology, Poland* The Application of Deep Convolutional Neural Networks in the Automated Diagnosis of Early Alzheimer Disease on Magnetic Resonance Images
10:20 - 10:40	Michał Ciach, *University of Warsaw, Poland* Estimation of Intensities of Reactions Triggered by Electron Transfer in Top-Down Mass Spectrometry
10:40 - 11:00	**Conference photo and coffee**
11:00 - 13:00	**Social swimming** *(Entrance to Uniejow Thermal baths is not included in the registration fee. Guests accommodated in Zamek and Dom Pracy Twórczej have free entrance to the Therms for 3 hours.)*
13:00 - 14:00	**Lunch**
14:00 - 15:40	**Session T2** Chair: Bartek Wilczyński
14:00 - 14:40	**Keynote speaker:** Maciej Komosiński, *Poznan University of Technology, Poland* Undirected Evolution of Simulated Stick Creatures
14:40 - 15:00	Paweł Błażej, *University of Wroclaw, Poland* Properties of Alternative Genetic Codes in Comparison with the Canonical Genetic Code and Theoretical Codon Assignments

15:00 - 15:20	Małgorzata Wnętrzak, *University of Wroclaw, Poland* The Application of Multiobjective Approach in the Optimization of the Genetic Code
15:20 - 15:40	Jarosław Paszek, *University of Warsaw, Poland* Algorithms for Genomic Duplication Models
15:40 - 16:00	**Coffee break**
16:00 - 17:10	**Session T3** Chair: Paweł Górecki
16:00 - 16:20	Maciej Błaszczyk, *University of Warsaw, Poland* Theoretical Methods for Structure Prediction of Proteins and Protein-Peptide Complexes
16:20 - 16:35	Melania Nowicka, *Poznan University of Technology, Poland* Codon Pair Bias Optimization in Synthetic Open Reading Frame Design
16:35 - 16:50	Anita Dudek, *University of Warsaw, Poland* How Good are Poisson-Boltzmann Calculations for Protein Hydration Free Energy
16:50 - 17:05	Aleksandra Świercz (on behalf of Piotr Żurkowski), *Poznan University of Technology, Poland* Graph Algorithms for DNA Sequencing - Origins, Current Models and the Future
17:30 - 18:30	**PTBI Convention**
19:00 - 22:00	**Conference dinner**

FRIDAY, September 29

09:00 - 10:40	**Session F1** Chair: Anna Gambin
09:00 - 09:40	**Keynote speaker:** Jacek Błażewicz, *Poznan University of Technology, Poland* Bioinformatics for RNA world theory
09:40 - 10:00	Robert Nowak, *Warsaw University of Technology, Poland* Detecting Genomic Rearrangements Using Markers
10:00 - 10:20	Maciej Miłostan, *Poznan University of Technology, Poland* Sharing of Life Sciences Linked Data - State of the Art and Challenges
10:20 - 10:40	Krzysztof Mnich, *University of Bialystok, Poland* Critical Line Algorithm for Combining Classifiers
10:40 - 10:50	**Closing remarks**
10:50 - 11:30	**Coffee and lunchbox**
11:00 - 12:30	**Session F2 (Workshop session)** Chair: Sebastian Kmiecik

LIST OF POSTERS

01	Jakub Bartoszewicz Predicting the Pathogenic Potential of Novel DNA Sequences Using Deep Learning
02	Michał Boniecki SimRNA: a Coarse-Grained Method for RNA Folding Simulations and 3D Structure Prediction
03	Juan F. Carrascoza Mayen On the Origins of Life: Theoretical Studies of Reactions Catalyzed by Montmorillonite on Atmospheric-Like Gases
04	Kaja Chmielewska Interleukin 18 Influence on the Cardiovascular System Modeled and Analyzed Using Petri Net-Based Approach
05	Maciej P. Ciemny CABS-dock Standalone Application for Protein-Peptide Docking with Large-Scale Flexibility of the Protein Receptor
06	Aleksandra E. Dawid SURPASS Low-Resolution Coarse-Grained Protein Modeling
07	Karolina Dawid PyMOL Plugin for Visualization and Analysis of Protein-Peptide and Protein-Protein Complex Structures
08	Witold Dyrka Where the Probabilistic Grammar Meets the Contact Map
09	Florian Heinke Concepts of Protein Energy Profiling: Deciphering Energy Fingerprints in Protein Structures

10	Paulina Hyży Ebola Virus Multialignment - Analysis and Visualization
11	Bogumił M.Konopka Sharing the First MinION 3rd Generation Sequencing Experience
12	Wojciech Lesiński Predicting Survivor in Neuroblastoma Based on RNA-Seq Data
13	Wojciech Łabaj Comparison of Tools for Mutation Detection Using Thyroid Cancer Genome Sequencing Data
14	Ania Macioszek Identifying Enrichment in Signal from DNA Sequencing Data Using Hidden Markov Models
15	Paweł Mackiewicz Costs of Amino Acid Replacement Can Be Minimized by Mutational Pressure in Bacterial Genomes
16	Paulina H. Marek Toward High-Resolution Prediction of Protein-Peptide Complex Structures
17	Justyna Mika Reannotation of VDJ Segments in Complementarity Determining Region 3 (CDR3) in Data from TCR Sequencing
18	Joanna A. Miśkiewicz Bioinformatics Study of Structural Patterns in Plant MicroRNA
19	Melania Nowicka An Answer Set Programming Approach to Optimal Design of Synthetic Cell Classifier Circuits

20	Marcin Pacholczyk EMQIT: a Machine Learning Approach for Energy Based PWM Matrix Quality Improvement
21	Marcin Pacholczyk Searching for Cancer Signatures Using Data Mining Techniques
22	Agata Perlińska Role of the Magnesium in a Knotted Methyltransferase
23	Aneta Polewko-Klim Identification of Informative Variables in Neuroblastoma Patients
24	Michał B. Ponczek Bioinformatics in Blood Coagulation System
25	Agnieszka Rybarczyk Tabu Search Algorithm for RNA Partial Degradation Problem
26	Katarzyna Rżosińska The Quantitative Model of the Process of Differentiation of Macrophages and Their Effects on Atherosclerosis Plaque Stability Based on Time Petri Nets
27	Jakub Wiedemann LCS-TA to Identify Similarity in Molecular Structures
28	Jakub Wojciechowski Contact Groups Improve Performance of DCA Contact Prediction
29	Marta Dudek Molecular Dynamics Simulations of Heterochiral RNA Complexes

ABSTRACTS

ORAL PRESENTATIONS

FORECASTING RESIDUE-RESIDUE CONTACT PREDICTION ACCURACY

Paweł P. Woźniak[1], Bogumił M. Konopka[1], Gert Vriend[2], Małgorzata Kotulska[1]

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

[2.] Centre for Molecular and Biomolecular Informatics, Radboud University Medical Centre, Nijmegen, The Netherlands

ABSTRACT

Most of today's best residue-residue contact prediction methods start with Direct Coupling Analysis (DCA) of correlated mutations in multiple sequence alignment (MSA). Despite high prediction accuracy of about 40% for the 100 strongest predicted contacts, it is an average over a large protein set. A user who works on a single protein, thus, will not know if contacts were predicted with either much higher or much lower accuracy than that 40%. This is especially a problem when the predictions are used to steer experimental research. We introduce a regression model that forecasts (with an error of only 7 percentage points) the accuracy of DCA-based residue-residue contact predictions for individual proteins. All tested models were trained for two DCA methods - gplmDCA and PSICOV using parameters that describe the MSA, the predicted secondary structure, the predicted solvent accessibility, and the contact prediction scores for the target protein.

GAPREPAIRER – REPAIR PROTEIN STRUCTURES AND THEIR TOPOLOGY

Aleksandra I. Jarmolińska[1,2], Michał Kadlof[1,3], Paweł Dąbrowski-Tumański[1,4], Joanna I. Sułkowska[1]

1. Centre of New Technologies, University of Warsaw, Poland
2. College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Warsaw, Poland
3. Faculty of Physics, University of Warsaw, Warsaw, Poland
4. Faculty of Chemistry, University of Warsaw, Warsaw, Poland

ABSTRACT

Protein structure is fundamental for its function. Topology is an important part of this structure. And topology can only be studied for an unbroken backbone. Thus properly filling in unresolved parts of protein's structure (found in >25% of PDB deposits) is instrumental for its in silico studies, e.g. molecular dynamics.

Non-trivial topologies in proteins include knots1, lassos2, and links3. Such folds can most easily be broken by a careless repair. Yet no current modeling tool includes the topology, potentially introducing erroneous, in a hard-to-notice way, folds.

GapRepairer is a server that fills this gap. It redefines homology to include topology, and analyzes the topology of both templates and final models. It provides an easy, but not basic, interface to the Modeller engine. GapRepairer, with a database of repaired structures is available at gaprepairer.cent.uw.edu.pl.

[1]JI Sulkowska et al PNAS(2012)
[2]W Niemyska et al Sci.Rep(2016)
[3]P Dabrowski et al PNAS(2017)

CABS-FLEX STANDALONE APPLICATION FOR FAST SIMULATIONS OF FLEXIBILITY OF GLOBULAR PROTEINS

Tymoteusz Oleniecki[1,3,4], Maciej P. Ciemny[1,2], Mateusz Kurciński[1], Maciej Błaszczyk[1], Andrzej Koliński[1], Sebastian Kmiecik[1]

1. Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Warsaw, Poland
2. Faculty of Physics, University of Warsaw, Warsaw, Poland
3. College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Warsaw, Poland

ABSTRACT

The conformational flexibility of protein structures is crucial for their functions. Simulations of protein flexibility remain computationally costly for most of protein systems using classical modeling tools. We present a new standalone version of our method – CABS-flex method. The method combines a highly efficient, coarse-grained approach with all-atom modeling methods. The CABS-flex predictions stays in accordance with MD simulations and NMR conformational ensembles. The CABS-flex method was also successfully used for efficient simulations of protein flexibility in predictions of protein-peptide complexes and protein aggregation properties. The standalone CABS-flex application allows for customization of the simulation parameters, handling large-sized systems and provides a flexible framework for result analysis. The standalone CABS-flex version is freely available at biocomp.chem.uw.edu.pl/CABSflexApp/ and server version at: biocomp.chem.uw.edu.pl/CABSflex/.

EQUANT: A WEB SERVICE FOR ENERGY-BASED MODEL QUALITY ASSESSMENT

Sebastian Bittrich, Dirk Labudde
University of Applied Sciences Mittweida, Mittweida, Germany

ABSTRACT

We present eQuant: a web service for protein structure quality assessment. PDB entries and even more so in silico modeled structures vary in quality and may contain delicate errors. eQuant employs a knowledge-based potential to discriminate between native respectively nonnative models and to spot discrepancies for particular residues.

Every amino acid has a certain propensity to be either hidden in the hydrophobic core of a protein or exposed to the polar solvent, which can be expressed as energy value according to the inverse Boltzmann law. For each residue, the total energy value can be computed by summing up the propensities of spatially neighbored residues in the 3D structure. Furthermore, such potentials of mean force have proven valuable for fold recognition and structure comparison.

The method was designed and trained on CASP data sets and is now continuously evaluated by the associated CAMEO project and is accessible at:
https://biosciences.hs-mittweida.de/equant/

IN SILICO METHOD FOR TRANSFORMATION OF DNA/RNA APTAMERS INTO PEPTIDE NUCLEIC ACID APTAMERS

Alexander Eisold, Dirk Labudde
University of Applied Sciences, Mittweida, Germany

ABSTRACT

Aptamers are short single-stranded oligonucleotides that fold into a complex three-dimensional structure in vivo, which can specifically bind different types of molecules. This specificity makes them suitable for detection of various compounds in different environments and allows their application as biosensors. Nevertheless, since nucleic acids are rapidly degraded by nucleases, they can be replaced by peptide nucleic acids (PNA), which are resistant to these enzymes and to proteases. In order to do this, a peptide backbone composed of N-(2-aminoethyl)-glycine units takes in PNA the place of the oligonucleotide sugar-phosphate backbone. PNA and DNA/RNA aptamers with identical base sequences have similar target-binding affinity. We propose an in silico method to transform known DNA/RNA aptamers into their PNA equivalents and to subsequently test their target specificity. This method will allow cost reductions during synthesis of PNA aptamers for different applications.

WATSON MEETS WATSON - HOW TODAY'S LIFE SCIENCE TECHNOLOGIES CAN SHAPE THE CRIME SCIENCES OF TOMORROW

Dirk Labudde
University of Applied Sciences Mittweida, Mittweida, Germany

ABSTRACT

Bioinformatics, the computer-based study of information and relations in biological systems, has emerged from rather simple tasks to visualizing, analyzing, classifying and simulating complex dynamic networks of hundreds or even thousands of biological entities. With the exponential growth of data provided by ever-faster experimental high-throughput techniques, bioinformatics aims at providing and developing novel computational tools to cope with these data.

Such 'information explosions' are not only present in life sciences; the growth of data produced in our daily life, which is mostly due to digitalized ways of communication and information exchange, behaves similarly. With respect to digital forensics, analyzing these large amounts of data and identifying significant pieces of information have become of great importance in current crime investigations. As a major starting point in our research, bioinformatics and digital forensics share a common aspect: looking for a needle in a haystack. As a matter of fact, bioinformatics approaches can be adapted to be of great help in finding evidence and missing links in confiscated data - a process that is still conducted manually in investigations to this date. In this talk, analogies and links between bioinformatics and digital forensics are elaborated.

Keynote speaker

MDFS - A SOFTWARE LIBRARY FOR MULTIDIMENSIONAL FEATURE SELECTION IN SEARCH FOR RELEVANT GENES

Radosław Piliszek[1], Krzysztof Mnich[1], Szymon Migacz[2], Paweł Tabaszewski[2], Aneta Polewko-Klim[3], Wojciech Lesiński[3], Witold Rudnicki[3]

1. Computational Centre, University of Bialystok, Poland
2. Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Warsaw, Poland
3. Institute of Informatics, University of Bialystok, Poland

ABSTRACT

Identification of informative variables is often the most important step of the dataset analysis but it tends to be performed using quick filtering procedures. Unfortunately, variables informative only in synergistic interactions with other variables are removed due to simple 1D filtering criteria.

We developed a software library devoted to identification of variables that carry information on the decision variable using multi-dimensional approach. It is an implementation of an algorithm based on conditional mutual information provided as an R library with computational engines in C++ and CUDA C. The CUDA C implementation is an efficient GPU-based implementation.

The application of the library in bioinformatics is demonstrated using real-world gene expression data for different types of leukemia. It is shown that multi-dimensional analysis is more sensitive than simple one-dimensional t-test and returns more important markers and genes in a reasonable amount of time.

POPULATION-SCALE DETECTION OF COPY NUMBER VARIATIONS IN ARABIDOPSIS THALIANA GENOME

Małgorzata Marszałek-Zeńczak, Agnieszka Żmieńko, Paweł Wojciechowski, Marek Figlerowicz
Institute of Bioorganic Chemistry, PAS, Poznan, Poland

ABSTRACT

The widespread presence of copy number variation (CNV) has been reported early in the first phase of Arabidopsis thaliana 1001Genomes Project. The lack of a comprehensive CNV map for this species currently limits the discovery of the contribution of individual CNVs to adaptive traits. We took advantage of the genomic data revealed by the 1001 Genomes Consortium to create such a map. Raw genomic reads for 1,135 accessions were downloaded from the repository. CNVs were called separately for each sample or for the entire population by adopting multiple approaches: read-depth, discordant paired-end read mappings, split-read and combined, following testing and optimization of the analysis pipeline. We are in the process of merging these to create a high-quality CNV map. Our map will complement the excellent resources on genomic, epigenomic and phenotypic diversity, recently provided by the 1001 Genomes Consortium thus enabling the studies of the CNV impact on plant evolution and adaptation.

NEW APPROACHES TO DETERMINE RNA PSEUDOKNOT ORDER

Tomasz Żok, Maciej Antczak, Mariusz Popenda, Michał Żurkowski, Ryszard W. Adamiak, Marta Szachniuk
Poznan University of Technology, Poznan, Poland

ABSTRACT

RNA molecules are very flexible and form complex folds. One such fold is called a pseudoknot. It is an intricate set of RNA interactions, however unlike real knot it would unfold to a linear representation when melted or pulled by force. In experimentally solved structures there are pseudoknots formed over other folds of that type which adds another layer of complexity. To correctly represent RNA secondary structure with this feature, one needs to optimize the decomposition steps which separate different pseudoknots from each other. Here we briefly summarize our previous heuristic algorithm and describe new approaches which result in a better transformation from 3D coordinates to a correctly stored secondary structure. This updated pipeline was also made available in our webserver http://rnapdbee.cs.put.poznan.pl.

This research was partially supported by National Science Centre, Poland (grant 2016/23/B/ST6/03931).

CLASSIFICATION AND PATTERN RECOGNITION IN BIOINFORMATICS BY PROTOTYPE-BASED MACHINE LEARNING MODELS

Thomas Villmann
University of Applied Sciences Mittweida, Germany

ABSTRACT

Modern bioinformatics has to deal with several kinds of data ranging from gene expression matrix data, spectral data and metabolic pathway graphs to structural information for protein folding in genomics to name just a few. Data analysis of those data comprises clustering, patter recognition, classification and regression. Thereby, difficulties may emerge due to the huge amount and complexity of data to be processed or, in contrast, regarding the sparsity of available data. Further data can be affected by noise, systematic disturbances like drifts or rotations.

The analysis of those data by machine learning approaches requires robust tools with good generalization abilities. One of the most successful approaches for supervised learning during the last years are deep neural networks (DNN) as advanced multi-layer perceptrons with very sophisticated gradient descent strategies and the possibility of unsupervised training of network modules [1]. Although exceedingly successful one disadvantage of DNNs despite the model complexity is their interpretability except in the application to image classification. Support vector machines (SVM) provide a powerful alternative for classification learning [2]. The support vectors, which are data themselves, determine the class

Keynote speaker

borders and optimization of SVM is based on convex optimization with powerful solvers. A drawback of SVM is the model complexity, which may increase extremely for complicate tasks leading to expansive training. Further, violation of data restrictions like positive definiteness for dissimilarities may lead to difficulties [3].

Prototype based models (PBM) in machine learning are well-known in machine learning since the 80s of the last century. They constitute robust and noise tolerant models combined with an intuitive interpretability. The common idea is to distribute reference vectors in data space – the so-called prototypes. Depending on the given task unsupervised or supervised models are known, the most prominent of which are the self-organizing map (SOM) and the neural gas (NG) for clustering and data visualization, and the learning vector quantization algorithms (LVQ) for classification learning [4]. The advantage of PBM is their inherently interpretability. For example, the aim of LVQ is to position the prototypes into the class distribution centers such that they can be taken as class-typical references. In SOM and NG, prototypes are distributed according to the data density whereas SOM realize a dimensionality reduction mapping, preserving neighborhood relation between data under certain conditions [5].

Thereby, the frequently mentioned gap of theoretical foundations for PBM was almost closed during the last years [6]. Nowadays gradient techniques can be applied to ensure mathematical correctness in treatment. Thereby, PBM can deal also with non-metric or kernel data violating basic properties like symmetry, Euclidicity or positive definiteness [7,8,9]. Related to bioinformatics, the application of correlation measures is of great interest [10]. Another benefit

Keynote speaker

in case of classification problems is the possibility of relevance (correlation) learning weighting automatically the data feature (or their correlation) regarding their importance for the classification task, which became one of the most exciting abilities [11,12]. Recent developments include tangent metrics or differential-geometric metrics in Grassmann manifolds to cope with variations in data or transfer learning 13,14,15]. The latter approach also provides the possibility of processing symbol sequences as demanded in genomics by utilization of generalized Hankel matrices [16].

A further virtue of modern LVQ is that instead of optimizing the simple classification error more sophisticated classification assessments can be optimized including precision and recall, the area under a ROC-curve or the widely applied F-measure [17,18]. These LVQ variants are of special interest in medicine and bioinformatics because in this area, as well as in case of imbalanced data during learning, the generally applied classification error is frequently misleading.

Comparable to SVM, LVQ can be modified to realize a border sensitive classification model [19]. Here the prototypes become sensitive particularly to the class borders in order to obtain a better class discrimination and detection of class borders instead of class prototypical reference vectors [20].

According to the current trend of deep learning as most powerful approach the question arises whether DNN should be preferred instead of prototype-based LVQ models. Clearly, for many applications, the deep architectures offer a superior approach if a sufficiently huge amount of training data is present or if the task is located in a prominent application area widely considered by deep learning methods like, for example, image recognition. However, in the case of sparsely available

Keynote speaker

for non-standard pattern recognition problems data LVQ approaches become a serious alternative frequently outperforming deep approaches. Further, pre-trained modules from deep network can be easily integrated also in LVQ architectures to benefit from their feature extraction capabilities. Otherwise, a deep network can be equipped with a LVQ-layer replacing the fully connected MLP-layer of a standard deep feedforward network in the last layer [21]. Thus, the resulting Deep-LVQ can profit from special deep models like, for example, from convolutional neural networks acting then as adaptive filters in LVQ [22]. Combination of Deep-LVQ with relevance learning leads to powerful flexible deep architecture keeping the interpretability of LVQ.

[1] I. Goodfellow, Y. Bengio, A. Courville. Deep Learning. MIT Press, 2016.
[2] B. Schölkopf, A. Smola. Learning with Kernels. MIT Press, 2002.
[3] D. Nebel, M. Kaden, A. Bohnsack, T. Villmann. Types of (dis-)similarities and adaptive mixtures thereof for improved classification learning. Neurocomputing, 2017.
[4] T. Kohonen. The Self-Organizing Map. Springer-Verlag, 1997.
[5] T. Villmann, R. Der, M. Hermann, T. Martinetz. Topology preservation in self-organizing feature maps: Exact definition and measurement. IEEE Transactions on Neural Networks, 8(2):256-266, 1997.
[6] M. Biehl, B. Hammer, T. Villmann. Prototype-based models in machine learning. Wiley Interdisciplinary Reviews: Cognitive Science,7:92-111, 2016.
[7] T. Villmann, S. Haase, M. Kaden. Kernelized vector quantization and gradient descent learning. Neurocomputing, 147:83-95, 2015.
[8] M. Kaden, M. Lange, D. Nebel, M. Riedel, T. Geweniger, T. Villmann. Aspects in classification learning - Review of recent developments in Learning Vector Quantization. Foundations of Computing and Decision Sciences, 39:79-105, 2014.
[9] D. Nebel, B. Hammer, K. Frohberg, T. Villmann. Median variants of learning vector quantization for learning of dissimilarity data. Neurocomputing, 169:295-305, 2015.
[10] M. Strickert, F.-M. Schleif, U. Seiffert, T. Villmann. Derivatives of Pearson correlation for gradient based analysis of biomedical data. Inteligencia Artificial - Revista Iberoamericana de Inteligencia Artificial, 37:37-44, 2008.

Keynote speaker

[11] B. Hammer, T. Villmann. Generalized relevance learning vector quantization. Neural Networks, 15(8-9):1059-1068, 2002.
[12] P. Schneider, B. Hammer, M. Biehl. Distance learning in discriminative vector quantization. Neural Computation, 21:2942-2969, 2009.
[13] S. Saralajew, T. Villmann. Transfer Learning in Classification based on Manifold Models and its Relation to Tangent Metric Learning. Proc. Int. Joint Conf. on Neural Networks (IJCNN), Anchorage, 1756-1765, IEEE Press, 2017.
[14] M. Kirby, C. Peterson. Visualizing Data sets on the Grassmannian using self-organizing maps. Proc. 12th Workshop on Self-Organizing Maps and Learning Vector Quantization (WSOM2017+), p. 32-37, IEEE Press, 2017.
[15] T. Villmann. Grassmann manifolds, Hankel matrices and tangent metric models in classification learning. Machine Learning Reports, 11:22-25, 2017.
[16] M. Mohammadi, M. Biehl, A. Villmann, T. Villmann. Sequence learning in unsupervised and supervised vector quantization using Hankel matrices. Proc. 16th Int. Conf. Artificial Intelligence and Soft Computing - ICAISC, Zakopane. L. Rutkowski et al. (Eds.), p. 131-142, Springer International Publishing, Switzerland, 2017.
[17] M. Kaden, W. Hermann, T. Villmann. Optimization of general statistical accuracy measures for classification based on learning vector quantization. Proc. Eur. Symp. on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2014). M. Verleysen (Ed.), p. 47-52, i6doc.com-Publisher, 2014.
[18] T. Villmann, M. Kaden, W. Hermann, M. Biehl. Learning vector quantization classifiers for ROC-optimization. Comp. Statistics, 1-22, 2016.
[19] M. Kaden, M. Riedel, W. Hermann, T. Villmann. Border-sensitive learning in generalized learning vector quantization: an alternative to support vector machines. Soft-Computing, 19(9):2423-2434,2015.
[20] T. Villmann, A. Bohnsack, M. Kaden. Can learning vector quantization be an alternative to SVM and deep learning? J. Artificial Intelligence and Soft Comp. Res., 7:65-81, 2017.
[21] T. Villmann, M. Biehl, A. Villmann, S. Saralajew. Fusion of deep learning architectures, multilayer feedforward networks and learning vector quantizers for deep classification learning. Proc. 12th Workshop on Self-Organizing Maps and Learning Vector Quantization (WSOM2017+), p. 248-255. IEEE Press, 2017.
[22] T. Neumann, S. Hellbach, M. Wacker. Deep Learning and LVQ: Some first Results in Image Classification. Machine Learning Reports, 11:16-17, 2017.

Keynote speaker

MINING FUNCTIONALLY CONSERVED BUILDING BLOCKS IN BIOLOGICAL MACROMOLECULES

Florian Kaiser, Dirk Labudde
University of Applied Sciences Mittweida, Mittweida, Germany

ABSTRACT

The biological function of proteins and nucleic acids, such as riboswitches or ribozymes, relies on the correct arrangement of small substructural units to catalyze substrates, bind ligands, or to preserve an ordered state. These molecular building blocks have evolved to retain similar interaction patterns and geometrical features to ensure functionality. While some of these patterns are reflected in recurring sequence motifs, evolutionarily remote proteins may only share a small set of similar structural motifs, which are not obvious on sequence level.

We utilize methods originated from data mining to identify structurally conserved, sequence separated, and family-specific building blocks in macromolecular structure data. This allows for the derivation of libraries of functionally conserved units to classify protein family association and investigate ancestry.

The presented workflow enables the automatic discovery of molecular building blocks without any a priori knowledge.

THE APPLICATION OF DEEP CONVOLUTIONAL NEURAL NETWORKS IN THE AUTOMATED DIAGNOSIS OF EARLY ALZHEIMER DISEASE ON MAGNETIC RESONANCE IMAGES

Aleksandra Suwalska, Franciszek Binczyk
Silesian University of Technology, Gliwice, Poland

ABSTRACT

Aim of the work was to design, implement and test the deep convolutional neural network as a tool for the automated detection of Alzheimer disease(AD)/mild cognitive impairment (MCI), on nuclear magnetic resonance medical images. The proposed model of convolutional neural network (CNN) contains of: input layer, convolutional layer of 16 filters (of 5x5 pixel size), max pooling (size 2x2 pixels) and classification layers. The activation function used was RELU. In the study EDSD dataset was used. It includes 493 DTI scans from patients with AD, MCI and Healthy Controls. The first results performed on a test set with pre trained CNN allows to differentiate an early AD from healthy with 70% accuracy (86% sensitivity). It was proven that deep neural networks could be successfully used for the automated image analysis in diagnosis of early AD. Proposed network will be extended by addition of additional hidden layers and more MR sequences (T1,T2). Work was financed by 02/010/BKM17/0083.

ESTIMATION OF INTENSITIES OF REACTIONS TRIGGERED BY ELECTRON TRANSFER IN TOP-DOWN MASS SPECTROMETRY

Michał A. Ciach, Mateusz Krzysztof Łącki, Błażej Miasojedow, Frederik Lermyte, Dirk Valkenborg, Frank Sobott, Anna Gambin
University of Warsaw, Warsaw, Poland

ABSTRACT

. Electron transfer dissociation (ETD) is a versatile technique used in mass spectrometry for high-throughput characterization of proteins. It consists of several competing reactions triggered by the transfer of electron from its anion source unto the sample cations. Relative quantities of the products of these reactions can be retrieved from mass spectra.

Here, we study these results from the perspective of the reaction kinetics. A formal mathematical model of the ETD is introduced and parametrized by intensities of the existing reactions. Also, we introduce a method to estimate the reaction intensities by solving a nonlinear optimisation problem. The presented method is proves highly robust to noise on in silico generated data. What is more, our model explains a considerable amount of experimental results gathered under various experimental settings

UNDIRECTED EVOLUTION OF SIMULATED STICK CREATURES

Maciej Komosiński
Poznan University of Technology, Poznan, Poland

ABSTRACT

The presentation will consist of three parts. The first part will discuss Artificial Life as a field of study: its definition, research interests and research directions. I the second part, I will introduce the Framsticks simulator

http://www.framsticks.com/

in particular, its simulation model and genetic representations. Examples of evolutionary optimization in Framsticks will be demonstrated. The third, final part, will concern evolutionary processes that are not directed by well-defined, static fitness functions. Instead, I will show how non-directed coevolution can be modeled in Framsticks, and what problems are encountered when evolving complex behaviors starting from the simplest artificial organism.

Keynote speaker

PROPERTIES OF ALTERNATIVE GENETIC CODES IN COMPARISON WITH THE CANONICAL GENETIC CODE AND THEORETICAL CODON ASSIGNMENTS

Paweł Błażej, Przemyslaw Gagat, Małgorzata Wnętrzak, Paweł Mackiewicz
University of Wroclaw, Wroclaw, Poland

ABSTRACT

It is commonly believed that the standard genetic code (SGC) is universal but many alternative genetic codes has been also discovered. The presence of these alternatives implies important questions about the evolutionary directions of the codes. Here, we assessed differences between the SGC and existing alternative codes according to costs of amino acid replacement based on their polarity. In addition, we tested the properties of all possible theoretical genetic code, which differed from the SGC in the fixed number of changes in assignments of codons to amino acids. We found that the substantial fraction of the theoretical codes minimized costs of amino acids replacement better than the SGC. Interestingly, many types of codon reassignments observed in the alternative codes are also responsible for the significant improvement of the fitness measure. These findings suggest potential evolutionary directions of alternative genetic codes.

THE APPLICATION OF MULTIOBJECTIVE APPROACH IN THE OPTIMIZATION OF THE GENETIC CODE

Małgorzata Wnętrzak, Paweł Błażej, Paweł Mackiewicz
University of Wrocław, Wroclaw, Poland

ABSTRACT

One of the theories about the standard genetic code origin postulates that the code evolved to minimize the effects of deleterious mutations and translational errors on the amino acid level. This process involved most probably many amino acid properties, often competing with each other. To study the possible adaptability of the genetic code, we applied the Strength Pareto Evolutionary Algorithm to perform 8-objective optimization based on the costs of changes in various amino acid properties resulting from all possible single point mutations in codons. Our results show that it is possible to find codes better optimized for every objective than the standard genetic code, but the latter is still definitely closer to the best optimized codes than to the worst optimized ones. It implies that some tendency to minimize the costs of amino acids replacements is present in the standard genetic code.

ALGORITHMS FOR GENOMIC DUPLICATION MODELS

Jarosław Paszek, Paweł Górecki
University of Warsaw, Warsaw, Poland

ABSTRACT

Discovering the location of gene duplications and multiple gene duplication episodes is a fundamental issue in evolutionary molecular biology. The idea is to map gene duplication events from a collection of gene family trees onto theirs corresponding species tree in such a way that the total number of multiple gene duplication episodes is minimized. Existing approaches vary in the two fundamental aspects: the choice of evolutionary scenarios that model allowed locations of duplications in the species tree, and the rules of clustering gene duplications from gene trees into a single multiple duplication event. We present mathematical foundations for general genomic duplication problems. We study the method of clustering called minimum episodes for several models of allowed evolutionary scenarios and propose efficient algorithms (e.g. first linear time algorithm for any interval model in which every gene duplication has an interval consisting of allowed locations in the species tree).

THEORETICAL METHODS FOR STRUCTURE PREDICTION OF PROTEINS AND PROTEIN-PEPTIDE COMPLEXES

Maciej Błaszczyk
Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Warsaw, Poland

ABSTRACT

The main purpose of my PhD thesis was to develop methods for structure prediction of proteins and protein-peptide complexes. The methods are based on multiscale modeling scheme merging highly effective coarse-grained CABS model with all-atom modeling tools. They have been made available as web servers: CABS-fold and CABS-dock.

The CABS-fold [1] web server is a tool for protein structure prediction from sequence only (de novo modeling) and also using alternative templates (consensus modeling). It is availabe at http://biocomp.chem.uw.edu.pl/CABSfold/

CABS-dock [2,3] web server provides an interface for modeling protein-peptide interactions using a highly efficient protocol for the flexible docking of peptides to proteins. While other docking algorithms require pre-defined localization of the binding site, CABS-dock does not require such knowledge. Given a protein receptor structure and a peptide sequence (and starting from random conformations and positions of the peptide), CABS-dock performs the simulation to search for the binding site allowing for full flexibility of the peptide and small fluctuations of the receptor backbone. CABS-dock is available at http://biocomp.chem.uw.edu.pl/CABSdock/

[1] Blaszczyk M, Jamroz M, Kmiecik S, Kolinski A. CABS-fold: Server for the de novo and consensus-based prediction of protein structure. Nucleic Acids Res. 2013;41: W406–11.
[2] 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. 2015;43: W419–24.
[3] Blaszczyk M, Kurcinski M, Kouza M, Wieteska L, Debinski A, Kolinski A, et al. Modeling of protein-peptide interactions using the CABS-dock web server for binding site search and flexible docking. Methods. 2016;93: 72–83.

CODON PAIR BIAS OPTIMIZATION IN SYNTHETIC OPEN READING FRAME DESIGN

Melania Nowicka
Poznan University of Technology, Poznan, Poland

ABSTRACT

Unrestricted manipulation of gene expression in heterologous host is one of the main objectives in synthetic biology. Increased expression may be achieved by an adaptation of codon composition in a gene to translational capabilities of a new host. Many studies reveal that previously used measures based on single codon arrangements are insufficient. Another, well-performing method for expression regulation is codon pair bias optimization. Unfortunately, precise gene expression control is a complex problem. Designing an open reading frame (ORF), one should consider other features affecting translation efficiency such as undesirable mRNA secondary structures, homopolymeric stretches, restriction enzymes sites or codon distribution. Available software does not meet several of those requirements. CodonComposer is an open source tool for codon pair bias optimization considering other above-mentioned features. Implemented genetic algorithms allow optimization of ORFs for applications such as industrial protein production or design of live attenuated virus vaccines.

HOW GOOD ARE POISSON-BOLTZMANN CALCULATIONS FOR PROTEIN HYDRATION FREE ENERGY

Anita Dudek, Piotr Setny
University of Warsaw, Warsaw, Poland

ABSTRACT

The stability of protein conformations is determined by the properties of their free energy landscapes. Although our attention usually focuses just on the details of protein molecular structures, we shouldn't forget that they naturally function in a dense aqueous environment. Water may shape proteins free energy landscape to no lesser degree than internal protein interactions.

In theoretical approaches, the role of water can be accounted for by conducting simulations in which a shell of water particles is explicitly present around the system of interest. Such simulations are, however, computationally demanding and their use in areas requiring long or numerous simulations such as protein folding or drug design studies is often prohibitive. An alternative is offered by so-called implicit solvent models, which treat solvent as a continuous medium, described by macroscopic physical parameters such as surface tension and dielectric constant. While well validated in the context of small molecules hydration, fidelity of such methods in the context of macromolecules with complex surface topographies and strong local electric fields has been not directly evaluated.

To precisely quantify solvent effects and evaluate the performance of implicit solvent models, we developed a protocol allowing for precise calculations of protein hydration free energy changes in explicit solvent simulations. Benefiting from the fact that free energy is a function of state and its changes do not depend on particular path between two endpoints, we consider nonphysical transformations between distinct protein conformations. Accompanying changes in free energy are then evaluated with thermodynamic integration method. The comparison of results with estimates obtained with Poisson-Boltzmann method allows the assessment of accuracy of this most widely used implicit solvent model. It turns out that Poisson-Boltzmann often gives results different from explicit solvent method. Our findings can be used to pinpoint the sources of problems faced by implicit solvent description of solutes having complex topologies and lead to its improvements.

GRAPH ALGORITHMS FOR DNA SEQUENCING - ORIGINS, CURRENT MODELS AND THE FUTURE

Piotr Żurkowski, Aleksandra Świercz
Poznan University of Technology, Poznan, Poland

ABSTRACT

The thesis focuses on the design and development of DNA de-novo assembly algorithm. The algorithm consists of two steps - construction of DNA overlap graph and the graph traversal. In order to generate a graph with the highest possible quality, a sophisticated heuristic algorithm has been combined with a complete alignment calculation - producing extraordinary graphs, albeit the computation time might be long.

Graph traversal is achieved by a novel algorithm, along with a number of quality improvements focusing various types of errors in the graph - including a fork detection algorithm, the most important phase regarding the quality of the contigs.

Multiple tests have been conducted on H. Sapiens and C. Elegans genomes and the results have been compared with state of the art assemblers. Both datasets showed a great performance of the proposed algorithm, which usually produced results of better quality than the other assemblers.

BIOINFORMATICS FOR RNA WORLD THEORY

Jacek Błażewicz
Poznan University of Technology, Poznan, Poland

ABSTRACT

According to some hypotheses, from a statistical perspective the origin of life seems to be a highly improbable event. Although there is no rigid definition of life itself, life as it is, is a fact. One of the most recognized hypotheses for the origins of life is the RNA world hypothesis. Laboratory experiments have been conducted to prove some of its assumptions. However, despite some successes in the "wet-lab", we are still far from a complete explanation. Bioinformatics, supported by biomathematics, provides perfect tools to model and test various scenarios of life's origins where wet-lab experiments cannot reflect the true complexity of the problem. Bioinformatics simulations of early pre-living systems may give us clues to the mechanisms of evolution. Whether or not this approach succeeds is still an open question. It seems likely that linking efforts and knowledge from various fields of science into a holistic bioinformatics perspective gives the opportunity to come one step closer to a solution to this question, which is one of the greatest mysteries of humanity. The talk illustrates some recent advancements in that area and points out possible directions for further research.

Keynote speaker

DETECTING GENOMIC REARRANGEMENTS USING MARKERS

Robert Nowak, Maciej Kulawik
Warsaw University of Technology, Warsaw, Poland

ABSTRACT

We present the application to detect large genomics rearrangements based on short sequences (markers). The algorithm develops the markers using the reference genome, the marker sequences are unique and its positions are uniformly spread. Then the markers are searched on one or many studied genomes, using few algorithms eg cross-correlation of signals using discrete Fourier transform. Finally the order of the markers is analyzed to report the rearrangements.

We develop the application using bioweb skeleton, C++, Python, PostgreSQL and JavaScript, three-layered architecture, the data layer and the calculation layer is deployed on server site. The end user needs only web browser. The application supports input files in FASTA and FASTQ, each user is able to run many independent calculating tasks, the output is GFF file and CIRCOS pictures.

We found rearrangements on few lines of Cucumis Sativus L. The demo server is available at http://antakya.ise.pw.edu.pl:9005.

SHARING OF LIFE SCIENCES LINKED DATA - STATE OF THE ART AND CHALLENGES

Maciej Miłostan
Poznan University of Technology, Poznan, Poland

ABSTRACT

Mining of biological data to discover knowledge relevant to disease, therapy, metabolic pathway rises a need to correlate and enrich data from heterogeneous sources (from a variety of experiments). It is not entirely automated process, and scientific questions raised by the researcher are its driving factor. The questions could come from various perspectives (the pathway, the molecule, the network) and determine the source of data. Discovery of the relationships among data is crucial to set hypothesis or confirm existing ones.

The problem of searching and linking of distributed life sciences data is not new. The prior efforts include definitions of ontologies and taxonomies, sharing interfaces based on RDF and SPARQL, the Life Sciences Linked Open Data Cloud, standardized formats, reasoning based on ontologies.

During the talk, an analysis of SOTA will be followed by an overview of challenges impacting sharing from the technological and non-technological point of view (e.g. GDPR).

CRITICAL LINE ALGORITHM FOR COMBINING CLASSIFIERS

Krzysztof Mnich, Witold Rudnicki
University of Białystok, Bialystok, Poland

ABSTRACT

Classification is a common machine learning task for biological data. It is used for medical diagnostics, predicting of protein properties, gene interactions investigating, metabolic pathway reconstruction etc. The key issue in classification algorithms is combining many weak, mutually dependent classifiers into the efficient classification function. Many approaches are used, like naive Bayes approach, linear models (e.g. LASSO) or numeric optimisation algorithms such as boosting.

We propose an efficient approach to combine classification functions. The only assumption about the elementary classifiers is the absence of strong synergistic interactions, that would reverse the result of classification.

We apply a linear approximation of the maximum likelihood, weighted Bayesian classifier. This leads to the well-known Non-negative Garrotte model, that can be solved exactly by Critical Line algorithm. The method has been tested on real, biological data.

ABSTRACTS

POSTERS

PREDICTING THE PATHOGENIC POTENTIAL OF NOVEL DNA SEQUENCES USING DEEP LEARNING

Jakub Bartoszewicz[1,2], Stefan Budach[2,3], Carlus Deneke[1,4], Robert Rentzsch[1], Annalisa Marsico[2,3], Bernhard Y. Renard[1]

1. Robert Koch Institute, Berlin, Germany
2. Free University of Berlin, Berlin, Germany
3. Max Planck Institute for Molecular Genetics, Berlin, Germany
4. Federal Institute for Risk Assessment, Berlin, Germany

ABSTRACT

Novel pathogens are expected to arise due to their fast-paced evolution or even genome engineering, and controlling them is needed to ensure public health, biosafety and biosecurity. Assessing the danger that unknown or artificial DNA sequences may pose is therefore crucial. However, traditional approaches are unfit for analysis of sequences highly divergent from known references, suggesting development of alternative tools using machine learning. As appropriate countermeasures must be deployed immediately, rapid assessment is key. This can be done by integration of pathogenicity prediction with a real-time Illumina read mapper, processing DNA reads while a sequencer is running. Here, deep neural networks are shown to perform better than the state-of-the-art approach and to capture features that would not be otherwise considered. Our results suggest that deep learning can be used to predict complex traits of whole organisms based on short, incomplete fragments of their DNA.

SIMRNA: A COARSE-GRAINED METHOD FOR RNA FOLDING SIMULATIONS AND 3D STRUCTURE PREDICTION

Michał Boniecki, Grzegorz Łach, Wayne K. Dawson, Konrad Tomala, Paweł Łukasz, Tomasz Sołtysiński, Kristian M. Rother, Janusz M. Bujnicki
International Institute of Molecular and Cell Biology, Warsaw, Poland

ABSTRACT

We developed a computational method for RNA folding simulations and 3D structure prediction, 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. In fact, those three atoms are used to calculate local coordinate system that allows for positioning of 3D grid - actual representation of the base. 3D grid contains information about interaction of the entire base moiety (not just 3 atoms).

Recent tests demonstrated that SimRNA is able to predict basic topologies of RNA molecules with sizes up to about 50 nucleotides, based on their sequences only, and larger molecules if supplied with appropriate distance restraints (secondary structure, pair-wise distance, and position restraints).

ON THE ORIGINS OF LIFE: THEORETICAL STUDIES OF REACTIONS CATALYZED BY MONTMORILLONITE ON ATMOSPHERIC-LIKE GASES

Juan F. Carrascoza Mayén[1], Natalia Szóstak[1], Jakub Rydzewski[2], Jacek Błażewicz[1,3], Wiesław Nowak[2]

[1.] Poznan University of Technology, Poznan, Poland
[2.] Nicolaus Copernicus University, Torun, Poland
[3.] Institute of Bioorganic Chemistry, PAS, Poznan, Poland

ABSTRACT

Preliminary results on Car-Parinello molecular dynamics of computational boxes containing elementary gases in contact with montmorillonite, a mineral clay commonly found on planet earth surface, suggest the formation of molecules crucial for the formation of life, such as formamide and others molecules containing the important carbon – carbon bond. These results are consistent with previous experimental findings and they suggest to be crucial to understand the further formation of polymeric compounds of life such as proteins and RNA that has been already described in other works. Our results are the first attempt to understand at quantum mechanical level a detailed mechanism of reaction and the potential roll deployed by the presence of mineral clays, all of this using pseudopotentials as described by the Car-Parinello dynamics theory.

INTERLEUKIN 18 INFLUENCE ON THE CARDIOVASCULAR SYSTEM MODELED AND ANALYZED USING PETRI NET-BASED APPROACH

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

[4.] Poznan University of Technology, Poznan, Poland
[5.] Poznan University of Medical Sciences, Poznan, Poland
[6.] Institute of Bioorganic Chemistry, PAS, Poznan, Poland

ABSTRACT

Interleukin 18 is one of the pro-inflammatory cytokines and it plays an important role in stimulation of natural killers and T cells to interferon gamma synthesis in response to activity of different pathogens. IL18 is recognized as very important regulator of innate and adaptive immune responses. Recent studies confirmed that IL18 can be produced by preadipocytes and adipocytes, reflecting the share of fatty tissue in regulation of inflammatory and metabolic processes, which can be engaged in development of atherosclerosis. IL18 can be helpful in prediction of cardiovascular events, however, there is no definitive evidence of an accurate IL18 action mechanism. It became the reason for using systems approach based on Petri nets for the analysis of the influence of IL18 on the cardiovascular system in the human body. The analysis of the model was based mainly on t-invariants. Research has been partially supported by the National Science Centre (Poland) grant No. 2012/07/B/ST6/01537.

CABS-DOCK STANDALONE APPLICATION FOR PROTEIN-PEPTIDE DOCKING WITH LARGE-SCALE FLEXIBILITY OF THE PROTEIN RECEPTOR

Maciej P. Ciemny[1,2], Tymoteusz Oleniecki[1,3,4], Mateusz Kurciński[1], Maciej Błaszczyk[1], Paulina H. Marek[1,5], Andrzej Koliński[1], Sebastian Kmiecik[1]

1. Faculty of Chemistry, University of Warsaw, Poland
2. Faculty of Physics, University of Warsaw, Poland
3. College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland
4. Mossakowski Medical Research Centre, PAS, Poland
5. Faculty of Chemistry, Warsaw University of Technology, Poland

ABSTRACT

Protein-peptide interactions may involve large conformational changes of the protein that are challenging to study experimentally or computationally. We present a new standalone application based on the CABS-dock method for protein-peptide docking. CABS-dock performs a blind search for a binding site combined with an on-the-fly folding of a flexible peptide, while the protein fluctuates around its input conformation. In CABS-dock, conformational flexibility of the protein receptor can be increased, as in MDM2/p53 complex modeled with full flexibility of the intrinsically disordered regions. The results matched well the experimental data and provided insights into the possible role of unstructured regions. The standalone CABS-dock application allows for customization and provides a flexible framework for result analysis. CABS-dock is available as a standalone application at biocomp.chem.uw.edu.pl/CABSdockApp and as a web server at biocomp.chem.uw.edu.pl/CABSdock.

SURPASS LOW-RESOLUTION COARSE-GRAINED PROTEIN MODELING

Aleksandra E. Dawid, Dominik Gront, Andrzej Kolinski
University of Warsaw, Warsaw, Poland

ABSTRACT

Coarse-grained modeling of biomolecules has a very important role in molecular biology. In this work we present a novel SURPASS (Single United Residue per Pre-Averaged Secondary Structure fragment) model of proteins that surpasses the sampling efficiency of existing coarse-grained models. The design of the model is unique and strongly supported by the statistical analysis of structural regularities characteristic for protein systems. Coarse-graining of protein chain structures assumes a single center of interactions per residue and accounts for pre-averaged effects of four adjacent residue fragments. Knowledge-based statistical potentials encode complex interaction patterns of these fragments. Using the Replica Exchange Monte Carlo sampling scheme and a generic version of the SURPASS force field we performed test simulations of a representative set of single-domain globular proteins.

PYMOL PLUGIN FOR VISUALIZATION AND ANALYSIS OF PROTEIN-PEPTIDE AND PROTEIN-PROTEIN COMPLEX STRUCTURES

Karolina Dawid[1,2], Maciej P. Ciemny[1,3], Tymoteusz Oleniecki[1,4,5], Mateusz Kurciński[1], Maciej Błaszczyk[1], Andrzej Koliński[1], Sebastian Kmiecik[1]

1. Faculty of Chemistry, University of Warsaw, Poland
2. Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland
3. Faculty of Physics, University of Warsaw, Poland
4. College of Inter-Faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Poland
5. Mossakowski Medical Research Centre, PAS, Poland

ABSTRACT

A clear molecular visualization of three-dimensional structures of protein-peptide and protein-protein complexes is a necessity in studying these systems. Such visualizations may be performed using feature-rich PyMOL molecular graphics software, however using PyMOL can be troublesome for many scientists inexperienced in programming. Our goal was to create a PyMOL plugin, which provides users, especially beginners and those without any programming background, with a simple analysis and visualization tool. Our plugin is designed to work with PDB files and provides support dedicated to the results produced with our protein-peptide docking method, CABS-dock. The plugin may be used in the graphical interface mode for regular use, or as a command-line tool.

WHERE THE PROBABILISTIC GRAMMAR MEETS THE CONTACT MAP

Witold Dyrka
Wroclaw University of Science and Technology, Wroclaw, Poland

ABSTRACT

Probabilistic context-free grammars (PCFG) were applied to detection of protein fragments, such as families of binding sites and amyloidogenic peptides. Our recent study demonstrated that derivation trees of PCFG could convey information of protein contacts [1]. However, standard learning approaches cannot guarantee that the inferred PCFG would generate trees consistent with the contact map of analyzed protein. With increasing availability of experimental and predicted contact maps, we aim at integrating this knowledge into the grammar inference process. Here we propose a form of PCFG appropriate for representing contacts between amino acids, introduce a notion of the parsing consistent with the contact map, and develop a suitable extension of the chart parser. Our method has potential to model wider classes of protein domains in comparison to approaches based on multiple sequence alignments [2].

[1] Dyrka et al. arxiv.org/abs/1611.10078
[2] Hopf et al. Nature Biotech 35:128 (2017)

CONCEPTS OF PROTEIN ENERGY PROFILING: DECIPHERING ENERGY FINGERPRINTS IN PROTEIN STRUCTURES

Florian Heinke, Dirk Labudde
University of Applied Sciences Mittweida, Mittweida, Germany

ABSTRACT

Knowledge-based potentials (KBPs) are widely applied in ab initio folding and docking simulations, structure refinement and assessment pipelines, as well as in predicting nativeness of protein models and complexes.

We here propose and discuss the theoretical concepts of protein energy profiling. Applying a coarse-grained KBP to structures of interest, per-residue energies and, based on that, sequences of energy values - we refer to as protein energy profiles (EPs) - are obtained. As an abstraction of both sequence and structure properties, analyses of EPs can be of interest for deciphering stabilizing and function-specific fingerprints.

We here discuss general structure-EP correlations and further propose methods for computing pairwise and multiple EP alignments as well as for deducing alignment scores, which allows to automatically unveil common energetic properties in entire EP sets and stabilizing features of functionally related proteins.

EBOLA VIRUS MULTIALIGNMENT - ANALYSIS AND VISUALIZATION

Paulina Hyży, Jakub Tyrek, Norbert Dojer
University of Warsaw, Warsaw, Poland

ABSTRACT

Multiple alignment is a rich source of knowledge. However, even a deep analysis is not enough to make this knowledge easily available for other researchers. To fully benefit from it one has to visualize the research results in a way convenient for others to consume.

As an example of this approach we will present multialignment of the Ebola virus. We extract the most relevant information, golden paths predictions and the sequences structure as a single model rendered by a web browser. The visual readability and computational efficiency is achieved by the usage of graph representation.

SHARING THE FIRST MINION 3RD GENERATION SEQUENCING EXPERIENCE

Bogumił M. Konopka, Agnieszka Rumińska, Samantha Filipow, Łukasz Łaczmański
Wroclaw University of Science and Technology, Wroclaw, Poland

ABSTRACT

Oxford Nanopore Technologies' MinION sequencer is a smartphone-sized sequencing machine. The sequencing method is based on measuring disturbances in current flow caused by DNA molecule passing through nanopores that are situated on a matrix of extremely sensitive current sensors. Based on the current disturbance pattern it is possible to identify the sequence of nucleotides that occlude the pore. Nanopore sequencing produces long reads that can be analyzed in real time. Thanks to its characteristics MinION can be successfully applied to structural variant calling, genome assembly or bacterial/viral identification. In this report we share our first MinION run on the standard Lambda phage kick-off sample (genome of 48 kb). We present the wet lab sample preparation procedure and data analysis, i.e. the analysis of the run meta-data, sequencing quality and time efficiency. We also investigate sequencing accuracy with respect to the reference genome.

PREDICTING SURVIVOR IN NEUROBLASTOMA BASED ON RNA-SEQ DATA

Wojciech Lesiński[1], Krzysztof Mnich[2], Aneta Polewko-Klim[1], Witold Rudnicki[1]

[1.] Institute of Informatics, University of Bialystok, Bialystok, Poland
[2.] Computational Centre, University of Bialystok, Bialystok, Poland

ABSTRACT

Neuroblastoma is an embryonal malignancy of the sympathetic nervous system. In this work we predict the patient's survivor based on RNA-seq data described in [1].The dataset contains RNA-seq expression profiles from 498 patients obtained using four different marker sets. To identify relevant variables we used standard t-test and our multidimensional feature selector (MDFS) [2], which is based on information theory. MDFS was able to identify variables involved in synergistic interactions. The Relevant variables was used to built random forest classification models.

Results: MDFS is more sensitive than t-test and variables that are found by this algorithm are more informative. Models built on most significant variables obtained from 2D MDFS give better prediction results than those from t-test.

[1] Zhang et al. Genome Biology (2015) 16:133
[2] K. Mnich, W. R. Rudnicki, arXiv:1705.05756

COMPARISON OF TOOLS FOR MUTATION DETECTION USING THYROID CANCER GENOME SEQUENCING DATA

Wojciech Łabaj, Andrzej Polański
Silesian University of Technology, Gliwice, Poland

ABSTRACT

The next generation DNA sequencing technology gives an opportunity for insight into the knowledge of tumor formation. One of the main goals of using NGS technology is somatic mutations detection.

Nowadays, there are many tools, which we are able to use for mutation detection. Some of them focus on mutation with high allelic fraction and filter out all mutations with low allelic fraction, which could be associated with either tumor heterogeneity or contamination by normal cells.

Next set of tools was designed to deal with this problem. However, scientists have a problem, which tools they have to use in particular cases. We have to take into account that each of the tools has a lot of parameters and discover mutations of particular nature, which is problematic.

Therefore, to face this problem and understand the obtained result, it was decided to compare four tools, which are the most known and commonly used for mutation discovery and show the relationships between them.

IDENTIFYING ENRICHMENT IN SIGNAL FROM DNA SEQUENCING DATA USING HIDDEN MARKOV MODELS

Ania Macioszek, Bartek Wilczyński
University of Warsaw, Warsaw, Poland

ABSTRACT

In modern biology experiments based on next generation sequencing play a key role. In many cases the result is a quantitative signal over a whole genome spanning billions of nucleotides. Frequently, the researchers want to identify regions of enrichment in signal corresponding to chromosomal locations associated with some biologically relevant property and distinguish them from noise. While there are many tools for analysing such data, they usually are either specialized to handle one particular type of experimental data or are general to the point of not being applicable to sequencing datasets. Here, we present a new approach, based on Hidden Markov Models; our tool allows integration of multiple sequencing tracks from replicate experiments as well as different experimental protocols.

COSTS OF AMINO ACID REPLACEMENT CAN BE MINIMIZED BY MUTATIONAL PRESSURE IN BACTERIAL GENOMES

Paweł Mackiewicz, Paweł Błażej, Dorota Mackiewicz, Małgorzata Grabińska, Małgorzata Wnętrzak
University of Wroclaw, Wroclaw, Poland

ABSTRACT

Mutations are usually regarded as a random process. Most of them are harmful for organisms but they are also essential for generation of genetic variation. Thus, we can expect that the mutational pressure is optimized to these two cases. To check the optimization of empirical mutational pressures, we compared bacterial nucleotide mutation matrices with their best possible alternatives. Using Evolutionary Multiobjective Optimization approach, we searched for the matrices that minimized or maximized costs of amino acid replacements resulted from differences in their physicochemical properties (e.g. hydropathy and polarity). The empirical mutational matrices appeared to have a tendency to minimize costs of amino acid replacements. It implies that bacterial mutational pressures can evolve to decrease consequences of amino acid substitutions. However, the optimization is not full, which enables generation a genetic variability necessary to adapt bacteria to changing environment.

TOWARD HIGH-RESOLUTION PREDICTION OF PROTEIN-PEPTIDE COMPLEX STRUCTURES

Paulina H. Marek[1,2], Maciej P. Ciemny[1,3], Mateusz Kurcinski[1], Maciej Błaszczyk[1], Andrzej Koliński[1], Sebastian Kmiecik[1]

1. Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Poland
2. Faculty of Chemistry, Warsaw University of Technology, Poland
3. Faculty of Physics, University of Warsaw, Poland

ABSTRACT

In recent years, peptides gained much interest in pharmaceutical research and development. Rational design of peptide therapeutics usually starts with structure characterization of a protein-peptide complex. In many cases it is difficult or impossible to use experimental approaches, thus reliable computational methods are needed. Practical applications require high-resolution predictions of sufficient accuracy for subsequent structure-activity relation analyses (i.e. studying the effect of in silico mutations). Here we present our results of using a combination of coarse-grained and high-resolution modeling methods that allow achieving high-resolution predictions (<1.5 angstroms) [1-3].

[1] Kurcinski, M., et al. Nucleic Acids Res 43, W419-424, (2015).
[2] Blaszczyk, M. et al. Methods 93, 72-83, (2016).
[3] Ciemny, M. P., et al. Methods Mol Biol 1561, 69-94, (2017).

REANNOTATION OF VDJ SEGMENTS IN COMPLEMENTARITY DETERMINING REGION 3 (CDR3) IN DATA FROM TCR SEQUENCING

Justyna Mika, Serge Candeias, Christophe Badie, Joanna Polańska
Silesian University of Technology, Gliwice, Poland

ABSTRACT

T cell receptors are one of the molecules responsible for big variety of immune responses. Their Complementarity Determining Regions (CDR) are created during VDJ recombination, a process consisting of selection and assembly of genome-coded V, D and J gene segments. During this process nucleotides might be inserted or deleted from joint regions, between V-D and D-J segments. Analysis of TCR sequencing data requires correct annotation of mentioned segments.

The study was performed on 16,694,800 sequences (517,087 unique) coming from deep sequencing of T cell receptors of 30 mice. Each sequence was automatically categorized by the selected V, D and J gene and described with the starting positions of every gene segment. Mistakes in gene classifications have been detected.

Own reannotation of gene segments based on correct open reading frames and knowledge about functionality status of segments improved gene classification, and was essential to perform reliable statistical inferring.

BIOINFORMATICS STUDY OF STRUCTURAL PATTERNS IN PLANT MICRORNA

Joanna A. Miśkiewicz[1], Marta Szachniuk[1, 2]

[1.] Institute of Computing Science & European Centre for Bioinformatics and Genomics, Poznan University of Technology

[2.] Institute of Bioorganic Chemistry, PAS, Poznan, Poland

ABSTRACT

A small non-coding molecule of microRNA (19-24nt) is related to both positive and negative aspects of various organisms lives. The amount of produced miRNA within an organism is highly correlated with key processes of human, plant or animal individuals, determines whether the system will work properly or defectively.

In plants, microRNA biogenesis requires DCL1 enzyme to perform the cleavages in pre-miRNA. This phase of miRNA maturation, recognition of miRNA in pre-miRNA structures by DCL1 in plants, remains an enigma in scientific world.

Herein we present a bioinformatic approach to discover specific motifs in the closest vicinity of plant microRNAs. We believe that in sequence or secondary structure of pre-miRNA occurs a motif/motifs which guide DCL1 enzyme to perform a cleavage right before the miRNA occurrence. To test our hypotheses we use known bioinformatic programs and develop scripts to analyze the data taken from a database of miRNA precursors – miRBase.

This research was partially supported by National Science Centre, Poland (grant 2016/23/B/ST6/03931).

AN ANSWER SET PROGRAMMING APPROACH TO OPTIMAL DESIGN OF SYNTHETIC CELL CLASSIFIER CIRCUITS

Melania Nowicka, Katinka Becker, Hannes Klarner, Heike Siebert
International Max Planck Research School for Computational Biology and Scientific Computing, Berlin, Germany

ABSTRACT

Cell classifiers are synthetic biological circuits capable of sensing endogenous molecular signals in a cell (inputs), interpreting them as type-specific signals, and triggering, e.g. cell apoptosis (output). Such selective cell targeting methods could provide more effective and non-toxic therapies for cancer patients. The inputs may be specified as binarized miRNA levels identifying the cell as healthy or diseased and combined into a Boolean expression classifying the cell state. The optimal circuit should accurately recognize diseased cells and consist of as few inputs as possible. Assembling the circuit in the laboratory poses additional constraints on feasible designs, e. g. specified gate types. Here we present an Answer Set Programming approach to the design of globally optimal classifiers and evaluation of the circuit response efficiency. The performance analysis shows that our method allows finding biologically plausible circuits taking only a few miRNAs as inputs.

EMQIT: A MACHINE LEARNING APPROACH FOR ENERGY BASED PWM MATRIX QUALITY IMPROVEMENT

Marcin Pacholczyk, Karolina Smolińska
Silesian University of Technology, Gliwice, Poland

ABSTRACT

We present EMQIT a modification to the approach introduced by Alamanova et al. and later implemented as 3DTF server. We observed that tuning of Boltzmann factor weights, used for conversion of calculated energies to nucleotide probabilities, has a significant impact on the quality of the associated PWM matrix.

Consequently, we proposed to use receiver operator characteristics curves and the 10-fold cross-validation to learn best weights using experimentally verified data from TRANSFAC database. We applied our method to data available for various TFs. Improved 3DTF matrices achieved significantly higher AUC values than the original 3DTF matrices (at least by 0.1) and, at the same time, detected notably more experimentally verified TFBSs. Matrices had comparable predictive capabilities. Moreover, improved PWMs achieve better results than matrices downloaded from 3DTF server. EMQIT is available online at
http://biosolvers.polsl.pl:3838/emqit

SEARCHING FOR CANCER SIGNATURES USING DATA MINING TECHNIQUES

Marcin Pacholczyk, Marta Micek
Silesian University of Technology, Gliwice, Poland

ABSTRACT

The goal of this work was to search for colorectal cancer signatures, consisting of somatic mutations, somatic gene copy number alterations (SCNAs) as well as abnormal expression levels. After acquiring mutation, SCNA and expression data from cBioPortal, frequent itemset mining was performed using basket analysis and apriori algorithm. We also performed survival analysis of colorectal cancer patients using the discovered signatures as differentiating factor for Kaplan-Meier curve comparison. Frequent itemset mining returned modifications of genes that can be regarded as potential colorectal cancer signatures or signatures of carcinogenic processes in general.

ROLE OF THE MAGNESIUM IN A KNOTTED METHYLTRANSFERASE

Agata P. Perlińska[1,2], Marcin Kałek[1], Joanna I. Sułkowska[1]

[1.] Centre of New Technologies, University of Warsaw, Poland

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

ABSTRACT

It was shown that Mg2+ is necessary for the proper functioning of one of the deeply knotted methyltransferases – TrmD protein[1]. Mg2+ was shown to be needed for catalysis, and should be present in the active site, close to the base that is methylated. Since the active site is a part of the knotted region[2], the necessity of the Mg2+ could be connected to the topology of the protein.

Using molecular dynamics simulations, we identified potential places for the Mg2+ to bind and studied how the presence of the ions influences the structure. Next, using DFT method we verified the locations of Mg2+ by calculating the energy barrier for the reaction with Mg2+ bound in each spot. We have found that TrmD has a preferred site for the magnesium, that could act both as a catalytic site and as a sensor for the Mg2+ concentration in the cell.

[1]Sakaguchi, et al. (2014). Chemistry biology 21(10), 1351-1360
[2]Christian, Sakaguchi, Perlinska, et al. (2016). Nature 23, 941-948

IDENTIFICATION OF INFORMATIVE VARIABLES IN NEUROBLASTOMA PATIENTS

Aneta Polewko-Klim, Krzysztof Mnich, Witold Rudnicki
Uniwersity of Bialystok, Bialystok, Poland

ABSTRACT

The genetic markers that carry information on difference in the clinical endpoints (survival and death) for 145 patients with neuroblastoma were examined using two datasets: gene expression profiles [1] and copy number profiles [2]. The identification of informative variables was performed using three methods: a standard t-test as well a 1D and 2D version of our MDFS [3] method based on information theory. There was significant difference between variables revealed by t-test MDFS for the gene expression aCGH dataset. The t-test, 1D-MDFS and 2D-MDFS find 5, 23 and 36 relevant genes, respectively. What is more, the Random Forest [4] models build using most relevant features obtained from MDFS have lower error than those obtained from t-test.

[1] Zhang, W. et al. Genome biol. 16, 2015.
[2] Theissen J. et al., GENE CHROMOSOME CANC, 53, 2014.
[3] Mnich, K. and Rudnicki W.R., arXiv:1705.05756, 2017.
[4] Breiman L. MACH LEARN, 45, 2001.

BIOINFORMATICS IN BLOOD COAGULATION SYSTEM

Michał B. Ponczek
University of Lodz, Lodz, Poland

ABSTRACT

Blood clotting involves complex processes which include chemical agents, proteins and non-protein biological elements, platelets in mammals, thrombocytes in other vertebrates, suspended in plasma. This sophisticated system has been studied biochemically for several decades. Such a complex system is an excellent subject for bioinformatics. Research in this field was launched in the 1960s by R.F. Doolittle starting from fibrinogen. The last two decades have enabled the sequencing of entire genomes of various organisms thanks to the development of new techniques together with automation of nucleic acid sequencing and brought the possibility of making information analysis within a reasonable period of time, due to growth in computing power. The aim of the poster is presenting the latest achievements at the interface between bioinformatics and biochemistry of hemostasis starting from sequence analysis through prediction of evolutionary associations to structure, function and drug design.

TABU SEARCH ALGORITHM FOR RNA PARTIAL DEGRADATION PROBLEM

Agnieszka Rybarczyk[1, 2], Alain Hertz, Marta Kasprzak[1, 2], Jacek Błażewicz[1, 2]

[1.] Poznan University of Technology, Poznan, Poland
[2.] Institute of Bioorganic Chemistry, PAS, Poznan, Poland

ABSTRACT

In the last few years, we have observed a great interest in the RNA research due to the discovery of the role that RNA molecules play in the biological systems. They do not only serve as a template in protein synthesis or as adaptors in translation process but also influence and are involved in the regulation of gene expression. It was demonstrated that most of them are produced from the larger molecules due to enzyme cleavage or spontaneous degradation.

In this work, we present our recent results concerning the RNA degradation process. In our studies, we used artificial RNA molecules designed according to the rules of degradation developed by Kierzek and co-workers. Base on the results of their degradation, we have proposed the formulation of the RNA Partial Degradation Problem (RNA PDP) and we have shown that the problem is strongly NP-complete. We would like to propose a new efficient heuristic approach, in which two tabu search algorithms cooperate. The algorithm can reconstruct a given RNA molecule, having as input the results of the biochemical analysis of its degradation, which possibly contain errors (false negatives or false positives). Results of the computational experiment, which prove the quality and usefulness of the proposed method, are presented.

THE QUANTITATIVE MODEL OF THE PROCESS OF DIFFERENTIATION OF MACROPHAGES AND THEIR EFFECTS ON ATHEROSCLEROSIS PLAQUE STABILITY BASED ON TIME PETRI NETS

Katarzyna Rżosińska[1], Dorota Formanowicz[2], Piotr Formanowicz[1, 3]

[1.] Poznan University of Technology, Poznan, Poland
[2.] Poznan University of Medical Sciences, Poznan, Poland
[3.] Institute of Bioorganic Chemistry, PAS, Poznan, Poland

ABSTRACT

In the development of atherosclerosis, alongside hyperholesterolemia and oxidative stress, the immunological processes play a crucial role. Homeostasis disturbances of tissue contributes to the development of local inflammation and growth of atherosclerosis plaque. Especially important is the process of macrophage differentiation. These cells can acquire pro- or anti-inflammatory abilities by taking the phenotype M1 or M2. M1 are particularly active in the acute phase of inflammation, and then subside for M2 macrophages in post-inflammatory phase. However, during the development of atherosclerosis, this pattern is disturbed. Time dependencies are very important in this process and it was taken into account in the proposed model. In this work a quantitative, time Petri net-based model of macrophage differentiation and their effect on the stability of atherosclerotic plaque is proposed. The model has been analyzed, with particular emphasis on structural analysis based on t-invariants.

LCS-TA TO IDENTIFY SIMILARITY IN MOLECULAR STRUCTURES

Jakub Wiedemann, Tomasz Żok, Maciej Miłostan, Marta Szachniuk
Poznan University of Technology, Poznan, Poland

ABSTRACT

Identification of common features and differences in biomolecule structures is an important task whose solution requires an involvement of bioinformatics methods. There is a necessity to develop and tune similarity measures to better analyse and evaluate structures, especially those predicted by computational approaches. Here, we present LCS-TA, a new method to detect local structural similarity. It finds the longest continuous segments in 3D structures folded in like manner. The folds are compared in torsion angle space and the measure of similarity is computed as the length of a segment.
This research was partially supported by National Science Centre, Poland (grant 2016/23/B/ST6/03931)

CONTACT GROUPS IMPROVE PERFORMANCE OF DCA CONTACT PREDICTION

Jakub Wojciechowski, Paweł P. Woźniak, Małgorzata Kotulska
Wrocław University of Science and Technology, Wroclaw, Poland

ABSTRACT

Information about residue-residue contacts can support the process of determination of protein three dimensional structure. The best contact prediction methods nowadays use Direct Coupling Analysis (DCA) based on correlated mutations. However, prediction of contacts is still challenging because even DCA methods can correctly predict on average only 20% out of 200 contacts with the highest score. We introduce a procedure that uses groups of contacts, which involves DCA score of a predicted contact and DCA scores between neighbouring amino acids, as an input for a neural network. The output of our network is then filtered using our other recently developed method which forecasts the prediction accuracy of DCA methods. As preliminary studies show, our procedure significantly improves precision of the best contact prediction methods.

MOLECULAR DYNAMICS SIMULATIONS OF HETEROCHIRAL RNA COMPLEXES

Marta Dudek, Joanna Trylska
CeNT, Warsaw, Poland

ABSTRACT

In my work I focus on identifying the interactions between natural ribonucleic acids (D-RNAs) and their mirror-reflected counterparts (L-RNAs). L-RNA differs from D-RNA only in sugar moiety that consists of β-L-ribose instead of β-D-ribose. Mirror-image RNA is a promising material for future drugs, capable to bind specifically selected biomolecules, with low toxicity and significant resistance to degradation in the organism. One of my goals was to perform pioneer MD simulations of heterochiral RNA complexes, which required modification of the AMBER ff10 force field to account for L-nucleotides. Additionally, simulations revealed many repeatable geometric patterns between non-canonically pairing strands that suggest possible heterochiral RNA motifs. The microscopic picture of L-RNA/D-RNA interface could help formulate the rules for designing L‑RNAs with desired affinity towards specific D-RNA.

AUTHOR INDEX