Session 4. Bioinformatics and Computational Biology

Lectures

L4.1

Retrogenes — genomic trash or treasure

Izabela Makałowska

Labolatory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
e-mail: Izabela Makałowska <izabel@amu.edu.pl>

It is estimated that the human genome contains as many as 8 000–15 000 copies of protein coding genes generated by reverse transcription. Although majority of them are in the state of a "relaxed" selection and remain "dormant", as they are lacking regulatory regions, many become functional. The evolutionary path of these functional retrogenes' is not uniform. In the course of the evolution they may undergo subfuctionalization and consequently share the function with their parent or develop a brand new function (neofunctionalisation). A good number of studies was recently performed to explore these unique sequences, yet our knowledge about retrogenes evolution and function is exceptionally limited. We performed a comprehensive analysis of 62 animal genomes to identify retrocopies of protein coding genes, study their evolution and decipher putative functions. Our analyses revealed that the fraction of expressed retrocopies is much higher than previously estimated and that retrocopies provide genetic material for new proteins, microRNA genes as well as trans natural antisense transcripts (trans-NAT). In addition, they may srve as microRNA sponges and, in case of nested retrocopies, possibly act as transcriptional interference factor and cause premature termination of host gene transcription. Moreover, our studies demonstrated remarkable differences in retrocopies evolution between placental mammals and other animals.

L4.2

Computational genome-wide cell-type-specific predictions of dimeric transcription factor complexes

Jerzy Tiuryn

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland
e-mail: Jerzy Tiuryn <tiuryn@mimuw.edu.pl>

The binding of transcription factors (TFs) to their specific motifs in genomic regulatory regions is commonly studied in isolation. However, in order to elucidate the mechanisms of transcriptional regulation, it is essential to determine which TFs bind DNA cooperatively as dimers or trimers, and to infer the precise nature of these interactions. We will discuss a novel method that allows for comprehensive prediction of TF dimers in a cell-type-specific way and on a genome-wide scale. Cell-type-specificity is achieved by utilizing DNase I hypersensitive sites that measure chromatin openness. Our results indicate that chromatin openness profiles are highly predictive of cell-type-specific TF-TF interactions. Moreover, they suggest that cooperative TF dimerization is a widespread phenomenon, and that most cell types are regulated by multiple TF complexes. We will also discuss a stand alone tool that allows to run locally the method of predicting cell-type specific dimers of TFs.

L4.3

Rational development of β2 adrenoceptor agonists capable to induce biased intracellular signalling

Krzysztof Jozwiak[1], Karolina Pajak[1], Anita Plazinska[1], Irving W. Wainer[2]

[1]Medical University of Lublin, Lublin, Poland; [2]Laboratory of Clinical Investigation, National Institute on Aging Intramural Research Program, Baltimore, MD, USA
e-mail: Krzysztof Jozwiak <krzysztof.jozwiak@am.umlub.pl>

The β2-adrenoceptor (β2-AR) is one of the best structurally and functionally characterized member of the G-protein coupled receptors family. The receptor has unique signaling properties: upon agonist induced activation couples to both G_s or G_i proteins and additionally can be involved in β-arrestin recruitment. In our medicinal chemistry projects a series of novel β2-AR agonists have been developed based on modification of fenoterol structure [1, 2]. The compounds exhibit very diverse array of functional efficacies in a number of β2-AR related assays *in vitro*. In one of the assays, the agonist induced cardiomyocyte contractility study very intriguing results have been observed where the small group of derivatives show a receptor activation pattern leading to coupling exclusively to G_s protein while most other derivatives activate the receptor to the forms capable to act *via* G_s and G_i coupling [3]. Structural analysis indicates that G_s selective group of derivatives have a specific substituent at 4' position capable to accept a hydrogen bond; subsequent molecular modeling simulations link this substituent with possibility of forming a hydrogen bond with Y308 residue of a β2-AR model. To verify that hypothesis Y308A mutant of β2-AR has been obtained and used to determine influence of the mutation on the binding affinities of derivatives. Indeed, G_s selective derivatives show their affinities to the Y308A mutant significantly reduced comparing to WT data, while the same mutation did not affect affinities for the group of derivatives eliciting both G_s and G_i signaling patterns. Further functional studies have indicated that the derivatives loose their G_s selective properties when assayed on the Y308F mutant of the receptor [4]. Therefore, a ligand hydrogen bond interaction with Y308 residue has been assigned as key event leading the receptor to the conformational active state which couples to G_s protein on a selective manner.
Currently, analogous experiments are performed using broader set of β2-AR mutants in order to track down other specific ligand – receptor interactions responsible for receptor activation biased to a specific intracellular signaling pathway.

References:

1. Jozwiak K, Khalid C *et al* (2007) *J Med Chem* **50**: 2903–2915.
2. Joziak K, Woo YH *et al* (2010) *Bioorg Med Chem* **18**: 728–736.
3. Woo YH, Wang TB *et al* (2009) *Mol Pharmacol* **75**: 158–165.
4. Woo YH, K. Jozwiak *et al* (2014) *J Biol Chem* in press.

L4.4

Understanding life: bioinformatics point of view

Jacek Błażewicz[1,2] , Natalia Szóstak[1]

[1]Institute of Computing Science, Poznań University of Technology, Poznań, Poland; [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland
e-mail: Jacek Blazewicz <jblazewicz@cs.put.poznan.pl>

From the statistical point of view the origin of life, according to older hypotheses, seems to be a highly improbable event. And, although there is no rigid definition of life itself, life as it is, is a fact. One of the most recognized new hypotheses for the origin of life is the RNA World hypothesis. Recently a lot of wet lab experiments have been conducted to prove some assumptions of it. However, regardless of the fact that some of them were successful, we are still far from the final explanation.
At the point where wet lab experiments still cannot reflect the true complexity of the problem, the bioinformatics seems to provide perfect tools to model and test various scenarios of the origin of life. Simulations of early preliving systems may give us clues to the mechanisms of evolution. Whether or not this approach succeeds, it is still an open question. However, it seems likely that linking efforts and knowledge from the various fields of science into bioinformatics holistic view have a chance to make us one step closer to a solution of this question being one of the greatest mysteries of mankind. The talk illustrates some recent advancements in that area and points out possible directions for further research.

Oral presentations

O4.1

The impact of sexual vs. asexual reproduction on transposon proliferation dynamics

Krzysztof Gogolewski[1], Michał Startek[2], Dariusz Grzebelus[3], Arnaud Le Rouzic[4], Anna Gambin[2]

[1]Faculty of Mathematics Informatics and Mechanics, University of Warsaw, Warsaw, Poland; [2]Institute of Informatics, University of Warsaw, Warsaw, Poland; [3]Department of Genetics, Plant Breeding and Seed Science, University of Agriculture in Krakow, Kraków, Poland; [4]Laboratoire Evolution, Génomes et Spéciation, Centre National de la Recherche Scientifique, France

e-mail: Krzysztof Gogolewski <k.gogolewski@student.uw.edu.pl>

It has recently come to general attention that transposable elements (TE) may have a significant influence on speciation and evolution of species. Thus understanding of transposon behavior and evolution seems crucial to deepening our knowledge on evolution of species.

The problem of transposon proliferation has been extensively studied from a theoretical point of view, and many models of this process have already been proposed. Most of such preexisting models are based on the concept of transposition-selection equilibrium (TSE) that is, the transposon counts within modeled populations are controlled by the struggle between transposons' selsh drive to duplicate regardless of the effect on their host, and the evolutionary drive to eliminate hosts with high transposon counts (which create a genetic burden resulting in lower viability of host organisms).

In TSE models, the transposons are in a state of equilibrium, while in nature they usually proliferate in short, intense bursts interspersed with long periods of relative calm. The burst periods are often related to a period of intense environmental stress, such as might be experienced when the species is colonizing a new ecological niche, or undergoing domestication by humans.

In most existing models of TE dynamics, the environment was considered constant. However, results of several experimental studies suggested that it should be viewed as one of the major factors when modeling the behavior of TEs in host genomes. In our approach, we analyze through stochastic simulations a model of TE dynamics that accounts for environmental pressure on host populations, i.e. the mutagenic activity of TEs, under the assumption of constant environmental conditions, is detrimental to host organisms. However, in a changing environment it may become beneficial because it allows the host population to adapt to the environment more efficiently. At the same time, stress lessens the usual negative effect of high mutability induced by heightened TE activity.

Our preliminary, computational, stochastic model dealt away with the concept of TSE, and instead it tracked the organisms' phenotypes (which were modified by transposition-induced mutations). The model was implemented for organisms undergoing asexual propagation. Here, we attempt to apply the same basic concept to investigate TE dynamics in sexually reproducing organisms.

Namely, we assume that each organism contains two, independent lists of transposons (two pairs of homologous chromosomes). In the course of reproduction it produces gametes containing a set of randomly chosen transposons from its genome, represented by the mentioned lists. Moreover, each transposon carries a unique value describing its contribution to the modification of the host phenotype. Finally, throughout the simulation we apply environmental stress, e.g. meteor impact or global warming scenarios as we did in the basic model. Under these new conditions we try to answer the question: What is the impact of sexual reproduction on proliferation of transposable elements as evolutionary helpers?

O4.2

p53 Regulatory module controlling cell cycle arrest and apoptosis in response to irradiation

Beata Hat[1], Marek Kochańczyk[1,*], Marta Bogdał[1], Tomasz Lipniacki[1,2]

[1]Institute of Fundamental Technological Research PAS, Warsaw, Poland; [2]Rice University, Houston, TX, USA

[*]presenting author

e-mail: Tomasz Lipniacki<tlipnia@ippt.pan.pl>

We construct a plausible model of p53 regulation in which cell fate decisions in response to irradiation are controlled by interlinked negative and positive feedback loops. Two primary negative feedbacks involve p53 arrester, its p53-responsive inhibitor, Mdm2, and p53-responsive ATM inhibitor, phosphatase Wip1. Existence of these two feedback loops enables oscillatory responses to DNA damage which can be terminated when DNA repair is completed (cell recovery) or due to switching to the state of high level of p53 killer (apoptosis). This bistable switch between (1) limit cycle oscillations characterized by high level of p53 arrester and very low level of p53 killer and (2) stable steady state characterized by very high level of p53 killer, arises due to two opposing positive feedback loops stabilizing respectively p53 arrester and p53 killer. The loop which stabilizes p53 arrester involves Wip1: its transcription is regulated by p53 arrester and Wip1 dephosphorylates p53 killer to p53 arrester. The loop which stabilizes p53 killer involves another phosphatase, PTEN, transcription of which is regulated by p53 killer and accumulation of which via Akt-dependent signaling leads to the sequestration of inhibitory Mdm2 in the cytoplasm, and in turn to stabilization of p53 killer at a high level. We analyzed the bifurcation structure of the model and found that, although quite complex, it is possibly the simplest one allowing for the coexistence of stable-period oscillations and a stable steady state "outside" of the limit cycle. In the oscillatory phase, p53 arrester triggers synthesis of p21, leading to the reversible cell cycle arrest, while in the high p53 killer steady state the activation of Bax transcription and inhibition of Akt lead to irreversible apoptosis. Analysis of the model demonstrated that cell fate decisions are controlled by expression levels of phosphatases Wip1 and PTEN, and the level of growth factors. These levels are highly variable between cancer cell lines – this can explain differences of their responses to irradiation.

O4.3

ClustSense: web service for sensitivity based parameters clustering

Karol Nienałtowski[1,2], Michał Komorowski[3], Anna Gambin[2,4]

[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland; [2]Mossakowski Medical Research Centre, Polish Academy of Sciences, Warsaw, Poland; [3]Division of Modelling in Biology and Medicine, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland; [4]Institute of Informatics, University of Warsaw, Warsaw, Poland
e-mail: Karol Nienałtowski <karol.nienaltowski@gmail.com>

Compared to engineering or physics problems dynamical models in quantitative biology typically exhibit substantially higher order of complexity. Therefore, the overall relationship between parameters and model dynamics is often prohibitively difficult to determine. Progress in developing tools to manipulate such models and so enable their efficient interplay with experiments has been slow, with methodological studies offering solutions significantly limited by model size.
In contrast, we have developed a general and efficient methodology dedicated to analyze the input - output relation of arbitrary large models of biochemical dynamics. Precisely, we quantify compensatory effects between model parameters. Our approach constitutes a natural mathematical language to precisely communicate and visualize parameter compensation phenomena that cannot be described with conventional tools and play an essential role in sensitivity analysis, parameter identifiability and experimental design.
We applied our methodology to reveal the role of parameters in the NF-$\varkappa$B signalling pathway. We surprisingly found that the relations between parameters correspond with the topology of the network. We described the consequences of this finding for model robustness, sensitivity and parameter identifiability. Importantly, our technique enables determination of non-identifiable parameters. We analyzed majority of experiments available in the literature on the NF-$\varkappa$B dynamics to verify which model parameters can be estimated. We have found that rich and heterogeneous experiments provide sufficient information to estimate only 22 out of the total 39 model parameters. This finding rises elementary questions regarding role of models with undetermined parameters and limits of parameter idenifiability of systems biology dynamical models.
Importantly, in order to make our method widely accessible we created a simple web application, which enables the analysis of biochemical dynamical models for users without quantitative expertise. Biochemical models are represented in the Systems Biology Markup Language. The computational procedures implemented in *Mathematica* and *R* environments are hidden from the user as a friendly interface is provided. Application is available on servers of Bioinformatics Laboratory of Mossakowski Medical Research Centre Polish Academy of Sciences.

O4.4

Novel approach for bi-clustering analysis in gene expression data

Paweł Foszner, Andrzej Polański

Institute of Informatics, Silesian University of Technology, Katowice, Poland
e-mail: Paweł Foszner <Pawel.Foszner@polsl.pl>

Bi-clustering is a data mining technique which allows simultaneous clustering of the rows and columns of a matrix. This technique belongs to the class of NP-hard problems, and was first presented by Morgan and Sonquist in 1963, than by Hartigan (1972), and by Mirkin in 1992. In the context of bioinformatics problems, the first to use this technique was Cheng and Church. They proposed bi-clustering of result from microarray experiments, by introducing the mean square residue in bi-cluster. Representative of modern algorithms can be QUBIC, introduced by the Guojun Li *et al.* They proposed very efficient algorithm for bi-clustering of matrix with discretized expression data. Authors use graph representation of data, and like Cheng and Church, also find bi-clusters with low mean square residue. The main problem of bi-clustering is that almost all of the existing algorithms specialize in a particular type of data. There are many types of data in a bi-clusters – constant values, plaid data, shift data, scaled data, and many others. Usually algorithms are dedicated to a particular type. Due to the fact that information about the bi-cluster is hidden, the researcher can never be sure what type of bi-clusters are present in the data. The bi-clustering analysis is usually time-consuming loop, in which researcher should choose the algorithm and set its parameters. This loop is repeated as long as he not find an algorithm that gives good quality results. This approach requires a lot of knowledge and experience, and most importantly – time.
The strategy of the performed research was oriented towards simplifying the analysis of bi-clustering to a pipeline as simple as possible: providing data on the input and getting the results on the output. The role of the user in this system is limited to the loading on the input data. However, it may also adjust the parameters used in the analysis. The key idea of proposed method is to computed large number of bi-clustering algorithms, each of which is specialized in different kinds of bi-clusters. Then, the results of these methods are combined into one. When paired, as the results we obtain set of sets of bi-clusters. Within a single set of bi-clusters, last step of algorithm is to connect bi-clusters composing it to a single one.
In this work we have shown that this approach gives satisfactory results regardless of data type of bi-clusters.

O4.5

RuleGO2 — a semi-interactive method and tool for generating most interesting logical rules for functional description of genes

Aleksandra Gruca[1], Marek Sikora[1,2], Łukasz Stypka[1]

[1]Institute of Informatics, Silesian University of Technology, Katowice, Poland; [2]Institute of Innovative Technologies EMAG, Poland
e-mail: Aleksandra Gruca <aleksandra.gruca@polsl.pl>

We present a new version of functional annotation tool - RuleGO [1] which allows describing groups of genes using combination of Gene Ontology terms [2]. Basic version of RuleGO Internet Service works the following way: the user submits set of genes she or he wants to analyze and reference set of genes. Based on GO terms composition of the both sets, the method generates all statistically significant GO terms combinations (so-called logical rules).
First version of RuleGO Internet Service allowed generating statistically significant combinations based on several rule quality assessment functions. As the method is looking for all possible combinations, the number of obtained rules is usually very large (around several thousands of rules). The rules are further filtered and the most interesting ones are presented to the expert.
However such approach may result in rejecting from the final output rule set some of rules that are interesting to the expert. To deal with that problem we created RuleGO2 which includes semi-automated rule generation algorithm which allows generating rules that more consistent with expert's preferences.
The interactive tool allows the expert to provide the set of the most interesting GO terms. For example, when analyzing cancer gene signature which allows distinguishing among different types of cancers, one can be specially interested in GO terms representing so-called hallmarks of cancer such as growth factors, apoptosis, angiogenesis, inflammation etc. [3]. The method tries to search for most interesting rules in three ways:
The user provides its own list of unfiltered rules and set of important GO terms. Obtained rules are then evaluated and analyzed by the filtration algorithm in such a way that rules containing important terms have the highest quality values.
The user provides gene signature, reference set and a list of important GO terms. The rules are generated in such a way that each rule include at least one of terms provided by the user.
The user provides gene signature, reference set and list of important GO terms. The rules are generated in such a way that each rule include at least one of the terms provided by the user. Then, if there are any genes from signature set which are not described by already generated rules, the algorithm tries to generate additional rules to obtain the highest possible coverage of the signature.
The tool is available at: www.rulego.polsl.pl

References:

1. Gruca A, Sikora M, Polański A (2011) RuleGO: a logical rules based tool for description of gene groups by means of Gene Ontology. *Nucleic Acids Res* **39** (suppl 2), W293-W301.
2. Ashburner *et al* (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* **25**: 25-29.
3. Hanahan D, Weinberg RA (2011) Hallmarks of Cancer: The Next Generation. *Cell* **144**: 646–674.

Acknowledgements

The work was supported by National Science Centre (DEC-2011/01/D/ST6/07007).

O4.6

Sorting signal targeting mRNA into hepatic extracellular vesicles

Natalia Szóstak[1], Felix Royo[3], Agnieszka Rybarczyk[1,2], Marta Szachniuk[1,2], Jacek Błażewicz[1,2], Antonio del Sol[4], Juan M. Falcon-Perez[3,5]

[1]Institute of Computing Science, Poznan University of Technology, Poznań, Poland; [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland; [3]Metabolomics Unit, CIC bioGUNE, CIBERehd, Bizkaia Technology Park, Derio, Bizkaia, Spain; [4]Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, Luxembourg; [5]IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
e-mail: Natalia Szóstak <nszostak@cs.put.poznan.pl>

Intercellular communication mediated by extracellular vesicles (EVs) has proved to play an important role in a growing number of biological processes including development, immunity, inflammation and tumour progression (Mathivanan *et al*., 2010; Ohno *et al*., 2012; Gutierrez-Vazquez *et al*., 2013). Among other molecules, mRNA seems to be one of the most interesting content of these vesicles. What we know is that mRNA localization depends on interactions between the cis-acting elements in the mRNA sequence, known as zipcodes, and trans-acting factors, the RNA-binding proteins. Here we present results of our research concerning zipcodes targeting mRNA into hepatic EVs. In order to perform efficient motif identification, we have combined two approaches: in silico and in vitro. During the in silico phase, as a first step, the data was analysed in order to obtain detailed information about the sequences. We have also checked whether mRNA sorting motifs, previously reported by other groups (Batagov *et al*., 2011; Bolukbasi *et al*., 2012) may act as cis-acting elements in hepatic cellular system. Negative correlation suggests that the mechanism of mRNA transport into EVs is tissue-specific. More importantly, based on bioinformatics tools, we have found 12 potential motifs, which may act as a zipcode for targeting mRNA into hepatic EVs. Secondary structure of these motifs have been predicted by mfold (Zuker, 2003), showing common folds for most of the candidate motifs. Additionally, miRNA-binding sites scan has been carried out using miRanda (John *et al*., 2004), detecting a number of miRNAs that could potentially bind the 12 selected motifs. This result supports the potential role of miRNAs in transporting mRNA into EVs. One of the putative zipcode, a 12-nt sequence included in a stem loop-forming region seemed to be particularly interesting. Its ability to target mRNA into EVs has been confirmed by a wet lab experiment. Taking into account that EVs serve as intercellular communicators, our results can have important therapeutics implications.

References:

Batagov AO, Kuznetsov VA, Kurochkin IV (2011) Identification of nucleotide patterns enriched in secreted RNAs as putative cis-acting elements targeting them to exosome nano-vesicles. *BMC Genomics* **12** Suppl 3: S18.
Bolukbasi MF, Mizrak A, Ozdener GB, Madlener S, Strobel T, Erkan EP, Fan JB, Breakefield XO, Saydam O (2012) miR-1289 and "Zipcode"-like Sequence Enrich mRNAs in Microvesicles. *Molecular Therapy Nucleic Acids* **1**: e10.
Gutierrez-Vazquez C, Villarroya-Beltri C, Mittelbrunn M, Sanchez-Madrid F (2013) Transfer of extracellular vesicles during immune cell-cell interactions. *Immunological Rev* **251**: 125–142.
John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS (2004) Human MicroRNA targets. *PLoS Biology* **2**: e363.
Mathivanan S, Ji H, Simpson RJ (2010) Exosomes: extracellular organelles important in intercellular communication. *J Proteomics* **73**: 1907–1920.
Ohno SI, Ishikawa A, Kuroda M (2012) Roles of exosomes and microvesicles in disease pathogenesis. *Adv Drug Delivery Rev* **2012**: 00249–00249.
Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**: 3406–3415.

O4.7

Signal peptide prediction using hidden Markov models

Michał Burdukiewicz[1], Piotr Sobczyk[2], Paweł Błażej[1], Paweł Mackiewicz[1]

[1]University of Wrocław, Department of Genomics, Wrocław, Poland; [2]Wrocław University of Technology, Institute of Mathematics and Computer Science, Wrocław, Poland
e-mail: Michał Burdukiewicz <pamac@smorfland.uni.wroc.pl>

Introduction: Secretory signal peptides are short N-terminal sequences directing a protein to endomembrane system and next to extracellular localization. Their amino acid composition allows to distinct three diverse N-terminal regions, the n-region (rich in positive amino acids) followed by the h-region (rich in hydrophobic amino acids) and the c-region (containing many polar and uncharged amino acids). The signal peptide is cut off from the mature part of protein by a specific cleavage site localized after the c-region (Nielsen & Krogh, 1998, *Proc Int Conf Intell Syst Mol Biol* **6**: 122–130). The secretory signal peptides are subjects of intensive researches because of their potential application in new drug development. Therefore, numerous methods of signal peptide prediction were elaborated using artificial neural networks (Petersen *et al*., 2011, *Nat Methods* **8**: 785–786), hidden Markov models (Käll *et al*., 2004, *J Mol Biol* **338**: 1027–1036), position weight matrices (Hiller *et al.,* 2004, *Nucl Acids Res* **32**: 375–379) and support vector machines (Cai *et al*., 2003, *Peptides* **24**: 159–161). However, they are still not perfect and do not work well on heterogenous data sets. The aim of this study was a new bioinformatics tool called signal.hmm for recognition of signal peptides based on hidden Markov models.
Materials and methods: Training and test protein sequences were taken from UniProt database. The sequences were filtered to remove atypical or poorly annotated records. Amino acids were aggregated into several physicochemical groups to generalize their properties and minimize dimensionality. The signal.hmm consists of two modules. The first one is a heuristic algorithm based on the current knowledge about the molecular structure and properties of secretory signal peptides. The algorithm extracts regions constituting the signal peptide from a given protein sequence. Such information is further used in the second module, which consists of two hidden Markov models recognizing proteins with and without the secretory signal peptide, separately. The implemented hidden Markov models have very flexible features. Instead of setting hard constrains on the length of regions typical of signal peptide, they calculate probabilities of staying in the given region. It allows the classifier to recognize correctly cases with abnormally short or long regions.
Results: The signal.hmm gives in outcome the probability of whether an analyzed sequence contains a signal peptide or not, as well as the position of possible signal peptide cleavage site. The AUC of the classifier is 0.97 and mean squared error of the predicted cleavage site is 6 positions.
Conclusions: Despite the project is in the initial phase, predictions of the signal.hmm are nearly as precise as its competitors. It is possible that the inherent extensibility guaranteed by the hidden Markov models will allow to reach eventually better results. The current problem concerns mostly the lack of precision in recognizing the end of cleavage peptide, which is caused by the similarity of the c-region and the beginning of mature protein.

O4.8

Modeling of large macromolecular complexes using hybrid approach

Joanna M. Kasprzak[1,2], Anna Czerwoniec[2], Mateusz Dobrychłop[2,1], Mateusz Koryciński[2], Janusz M. Bujnicki[1,2]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland; [2]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
e-mail: joanna kasprzak <jkasp@amu.edu.pl>

One of the major challenges in structural biology is to determine the structures of macromolecular complexes and to understand their function and mechanism of action. However, structural characterization of macromolecular assemblies is very difficult. A hybrid computational approach is required that will be able to incorporate spatial information from a variety of experimental methods into modeling procedure.
Thus far, we developed PyRy3D, a method for building low-resolution models of large macromolecular complexes. The components (proteins, nucleic acids and any other type of physical objects including e.g. solid surfaces) can be represented as rigid bodies (e.g. based on atomic coordinates of structures determined experimentally or modeled computationally) or as flexible shapes (e.g. for parts, whose structure is dynamic or unknown). The model building procedure applies a Monte Carlo approach to sample the space of solutions. Spatial restraints are used to define components interacting with each other, and a simple scoring function is applied to pack them tightly into contours of the entire complex (e.g. cryoEM density maps). This approach enables the construction of low-resolution models even for very large macromolecular complexes with components of unknown 3D structure, such as human mitochondrial RNA polymerase gamma.
We carried out a structural bioinformatics analysis of large macromolecular complexes such as: editosome, human Splicing Factor 3b (SF3b) and CASCADE, predicted disordered and ordered regions and modeled the structures of individual domains. Then we used PyRy3D to generate ensembles of models that fulfill restraints satisfying known experimental data. Such an approach is the first step to understand mechanism of actions for these enzymes and might help in structure determination with experimental methods.

O4.9

Binding of hydrophobic ligands to G-protein-coupled receptors

Jakub Jakowiecki[1], Dorota Latek[1], Shuguang Yuan[2], Sławomir Filipek[1]

[1]Faculty of Chemistry & Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland; [2]École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
e-mail: Sławomir Filipek <sfilipek@chem.uw.edu.pl>

G-protein-coupled receptors (GPCRs) form a very large and important family of membranous signaling proteins. These receptors are pharmacological targets for 30–50% of currently used drugs. More than 800 human GPCRs allow cells to recognize diverse extracellular stimuli and transduce the signals across the plasma membrane to activate G proteins which pass the signal to the effector proteins. Recently, it was found that arrestin is also mediating the signaling, independently of G protein [1]. Choice of a mediator and also a degree of increase or decrease of receptor activation, since GPCRs are partially activated even without a ligand, is regulated by binding of appropriate ligands. The ligands are traditionally divided into three classes: agonists – increase activation, antagonists – block the receptor and do not change activity, and inverse agonists which decrease activation. However, with two mediating proteins such classification is too simplistic. The hydrophobic ligands of GPCRs are not so common among other ligands of these proteins since their interior is rather hydrophilic with many water molecules participating in internal hydrogen bond network. Rearrangement of this network by ligand binding and action of microswitches [2] leads to movement of receptor helices and binding of mediating proteins. We investigate how the hydrophobic ligands can enter the GPCR binding site and which are their binding modes. The analysis is based on crystallographic structures with hydrophobic ligands and also after docking procedures, as for sphingosine-1-phosphate (S1P1) receptor [3] as well as based on homology models of GPCRs [4], as for cannabinoid receptors [5].

References:

1. Latek D, Modzelewska A, Trzaskowski B, Palczewski K, Filipek S (2012) *Acta Biochim Pol* **59**: 515–529.
2. Trzaskowski B *et al.* (2012) *Curr Med Chem* **19**: 10901109.
3. Yuan S, Wu R, Latek D, Trzaskowski B, Filipek S (2013) *PLoS Comput Biol* **9**: e1003261.
4. Latek D, Pasznik P, Carlomagno T, Filipek S (2013) *PLoS ONE* 8: e56742.
5. Latek D *et al.* (2011) *J Mol Model* **17**: 23532366.

O4.10

Molecular dynamics simulation studies of human cystatin C oligomerization

Magdalena Murawska[1], Sylwia Rodziewicz-Motowidło[2], Maciej Kozak[1]

[1]Department of Physics, Adam Mickiewicz University, Poznań, Poland; [2]Department of Chemistry, University of Gdańsk, Gdańsk, Poland
e-mail: Magdalena Murawska <murawska@amu.edu.pl>

Molecular dynamics simulations become a useful tool, enabling build of realistic atomistic models of biological systems and observation of their behavior over time, depending on the different calculation parameters (force field, temperature, pressure, etc).
Human cystatin C (HCC) is an inhibitor of cysteine proteases (papain, cathepsins B, H, K, L and S) and C13 family – legumain [1]. In medical diagnostics Cystatin C is used as an important marker of kidney function. For HCC was observed the tendency to oligomerization *via* the three-dimensional exchange of domain, called "domain swapping" [1, 2], which leads to aggregation of protein in the brain arteries of older people and cause amyloid angiopathy. Rarely, a naturally occurring mutation HCC (Leu68Gln) results in massive amyloidosis, cerebral hemorrhage and ultimately to death at a young age. Probably HCC is also involved in the formation of pathological amyloid fibrils in the brain of patients with Alzheimer's disease.
The aim of this study was an attempt to create models of various HCC oligomers using molecular dynamics calculations, based on the information provided by TEM and AFM microscopy. Obtained micro-images revealed that HCC oligomers have donut-like, planar morphology. The molecular dynamic simulations were performed using AMBER program package and several structural models of HCC oligomers were created. Models of HCC trimers were also compared with low-resolution structure of trimeric form of human cystatin C in solution, which was restored by a computer simulation in program DAMMIN [3]. The X-ray scattering data was obtained using synchrotron radiation and SAXS camera (beam line BL911-4 [4], MAXII storage ring of the MAX-Lab Lund, Sweden; λ= 0.091 nm).

References:

1. Janowski R *et al* (2001) *Nat Struct Biol* **8**: 316–320.
2. Wahlbom M *et al* (2007) *J Biol Chem* **282**: 18318–18326.
3. Svergun DI *et al* (1999) *Biophys J* **77**: 2879–2886.
4. Mammen CB *et al* (2002) *Acta Physica Pol* A **101**: 595–602.

Acknowledgements

The present study was supported by HARMONIA3 grant (Project No. DEC-2012/06/M/ST4/00036) from the National Science Centre, Poland.

O4.11

SimRNA: a program for RNA 3D structure prediction and folding simulations

Michał J. Boniecki[1], Grzegorz Łach[1], Konrad Tomala[1], Tomasz Sołtysiński[1], Paweł Łukasz[1], Kristian Rother[1,2], Janusz M. Bujnicki[1,2]

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

e-mail: Michał Boniecki <mboni@genesilico.pl>

The molecules of the ribonucleic acid (RNA) perform a variety of vital roles in all living cells. Their biological function depends on their structure and dynamics, both of which are difficult to experimentally determine, but can be theoretically inferred based on the RNA sequence. We have developed a computational method for molecular simulations of RNA, 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. Despite the bases being represented by only three atoms, other atoms can be implicitly taken into account in terms of the excluded volume. All base-base interactions were modeled using discrete three-dimensional grids built on local systems of coordinates.

All terms of the energy function used were derived from a manually curated database of crystal RNA structures, as a statistical potential. Sampling of the conformational space was accomplished by the use of the asymmetric Metropolis algorithm coupled with a dedicated set of moves. The algorithm was embedded in either a simulated annealing or replica exchange Monte Carlo method. 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. The user can specify various types of restraints, including restraints on secondary structure, distance and position.

SimRNA can be used for systems composed of several chains of RNA. It is also able to fold/refine structures with irregular (non-helical) geometry of the backbone (RNA pseudo-knots, coaxial stacking, bulges, etc.). As SimRNA is based on folding simulations, it also allows for examining folding pathways, getting an approximate view of the energy landscapes, and investigating of the thermodynamics of RNA systems.

O4.12

SimRNA-design: a computational method for designing RNA sequences that fold into desired 3D structures

Krzysztof Formanowicz[1, 2], Grzegorz Łach[1], Michał Boniecki[1], Janusz M. Bujnicki[1,2]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland; [2]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

e-mail: Krzysztof Formanowicz <kformanowicz@genesilico.pl>

Ribonucleic acids (RNAs) are a large group of biological molecules that play vital roles in many biological processes, such as coding genetic information, and catalyzing chemical reactions. They also perform regulatory functions. While RNA tertiary structure prediction is a classic problem in structural bioinformatics, the inverse problem of designing a sequence folding into a desired three-dimensional structure appears even more challenging. Artificial RNA sequences that form specific tertiary structures may have many potential biomedical and synthetic biology applications.

SimRNA-design is a computational method for RNA design based on the SimRNA simulation program. Like SimRNA itself it uses a coarse grained representation of RNA molecule, and statistical potential derived from large set of known RNA structures. Sampling of possible nucleotide sequences and structures is done using asymmetric Metropolis algorithm coupled with a dedicated set of spatial moves and sequence modifications. The simulation can be done using either simulated annealing or replica exchange Monte Carlo method. The user can specify various types of restraints, including geometry (distances and positions), secondary structure, sequence, and nucleotide composition restraints.

The output of SimRNA-design is a RNA sequence which has lowest free energy when folded into given desired tertiary structure, and which also is not likely to fold in any other way.

O4.13

Prediction of functional DNA elements from histone modifications

Joanna Giemza[1,2], Bartek Wilczyński[1]

[1]Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Warsaw, Poland; [2]Division of Genetics & Molecular Medicine, King's College London, United Kingdom
e-mail: Joanna Giemza <joanna.giemza@kcl.ac.uk>

Histones, especially their N-terminal tails, are subject to a large number of post-translational modifications. In the last few years, thanks to high-throughput experimental methods, particularly CHIP-chip and CHIP-Seq, remarkable progress has been made in the characterization of histone modifications. The link between histone modifications and transcription has been particularly intensively studied. It has been found [1, 2] that some individual modifications can be associated with transcriptional activation or repression. For instance, H3K4me3 is enriched in promoters and H3K36me3 in transcribed regions of active genes. Furthermore, it has been shown, based on genome-wide CHIP-Seq data for 38 histone modifications and one histone variant from human CD4+T-cells, that histone modification levels are predictive of gene expression levels [3].
We analysed the same dataset, but addressing different problem: prediction of functional DNA elements, such as promoters, exons and introns from histone modifications. We compared classifiers based on Bayesian networks and random forests. In the first case, we consider a bipartite Bayesian Network between classification attributes (histone modifications) and predicted classes (binary indicators of functional elements). BNFinder software [4] provides the optimal topology of the network, performing feature selection simultaneously. Additionally, we analyze impact of preprocessing on classification quality.
Our results from T CD4+ human cell line as well as fruit fly embryo indicate that it is possible to accurately predict major functional annotations in their active state, while inactive elements seem to be difficult to distinguish from intergenic regions. While random forest classifiers provide overall better accuracy, the Bayesian models are more useful in selecting the few most informative modifications.

References:

1. Artem Barski *et al* (2007) *Cell* **129**: 823–837.
2. Zhibin Wang *et al* (2008) *Nature Genetics* **40**: 897–903.
3. Rosa Karlić *et al* (2010) *PNAS* **107**: 2926-2931.
4. Bartek Wilczyński, Norbert Dojer (2009) *Bioinformatics* **25**: 286–287.

Acknowledgements

This work was supported by the Foundation for Polish Science within Homing Plus programme co-financed by the European Union - European Regional Development Fund.

O4.14

Inhibition of bacterial translation by Peptide Nucleic Acid oligomers targeting the ribosomal RNA

Anna Górska, Joanna Trylska

Centre of New Technologies, University of Warsaw, Warsaw, Poland
e-mail: Anna Górska <agorska@cent.uw.edu.pl>

We have been using short modified oligonucleotides, such as Peptide Nucleic Acids (PNA), to target in a sequence-specific manner the functional sites of bacterial ribosomal RNA (rRNA) andinhibit bacterial translation. To monitor the inhibitory efficiency of these oligomers, we have been using bacterial cell-free extracts. However, these experiments are difficult, time consuming andcostly, so advance computer-aided oligomer design to predict the inhibitory effectiveness is needed. For this reason I have been developing various software tools to evaluate in silico the effectiveness ofPNA sequences in targeting a particular region in rRNA. I considered three main features that determine the usefulness of a candidate PNA oligomer: ecological-range of impact, functionality of the targeted RNA region, and predicted capability of efficient hybridization. By ecological range I define a group of species that are susceptible to a particular PNA sequence. To rank the bacterial strains by the probability of its rRNA being targeted by a certainPNA oligomer I have developed a multistage bioinformatics pipeline. Data concerning the functionality of the region were derived from the crystal structures of the ribosomes in complexes with various ligands such as antibiotics or protein factors. Literature data on the mutagenesis analyses were also incorporated into the algorithm. The capability of efficient binding is the most complicated part of the score because apart from the obvious sequence-specificity it depends on many features of rRNA including its secondary and tertiarystructures, accessibility of a particular RNA fragment, and most importantly its native-state dynamics. I have also considered and evaluated the opening energy of targeted rRNA or its purine content (since PNA oligomers prefer homopurie regions). To assess the flexibility of rRNA in the ribosome context I performed all-atom molecular dynamics simulations of the small ribosomal subunit.The main difficulty in developing such a tool is integrating various types of information and selecting reliable parameters to score a particular PNA sequence. Due to insufficient amount of experimental data, the predictive power of the designed algorithm and software is being assessed in our lab by *in vitro* experiments that test the inhibition of bacterial translation by PNA oligomers in *E. coli* cell-free transcription/translation system.

References:

1. Nielsen P (1999) Peptide nucleic acid. A molecule with two identities. *Accounts Chem Res* **32**: 624–630.

2. Phillips J *et al* (2005) Scalable molecular dynamics with NAMD. *J Comp Chem* **26**: 1781–1802.

3. Zhang P (2010) Production of biocommodities and bioelectricity by cell-free synthetic enzymatic pathway biotransformations: challenges and opportunities. *Biotechnology and Bioengineering* **105**: 663–677.

O4.15

MetaMisTher: a machine learning meta-predictor of the influence of missense mutations on thermodynamical protein stability

Witold Januszewski[1,2], Łukasz Kozłowski[2], Marcin Magnus[2], Tymon Rubel[3], Janusz M. Bujnicki[2,4]

[1]Institute of Biology III, Albert Ludwig University of Freiburg, Germany; [2]International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland; [3]Institute of Radioelectronics, Warsaw University of Technology, Warsaw, Poland; [4]Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University of Poznan, Poznań, Poland

e-mail: Witold Januszewski <witold.januszewski@biologie.uni-freiburg.de>

Missense mutations, also known as nsSNPs (nonsynonymous Single Nucleotide Polymorphisms) are point mutations in DNA regions leading to a single amino acid substitution in the primary protein structure. The difference between free Gibbs energy of wild type and mutant protein ($\Delta(\Delta G)$) is the indicator of thermodynamical protein stability. Experimental techniques such as differential scanning calorimetry (DSC), fluorimetry and electrophoresis have been used to link missense mutations with certain free energy changes. Henceforth, theoretical predictions of postmutational protein stability can be validated and benchmarked easily. Empirically inforced constraints for $\Delta(\Delta G)$ were suggested by Khan: $\Delta(\Delta G)$ within the range {-0.5;0.5 kcal/mol} signifies neutral mutation, above it – destabilizing and below – stabilizing [1]. Although a particular missense mutation cannot be linked to a disease deterministically, several studies [2] show that pathogenicity predictions of missense mutations are also possible.

The main objective of the MetaMisTher project is to create an independent and intuitive metapredictor of the influence of missense mutations on thermodynamical protein stability. Another objective is to use as little biological expert data as possible for the task, thus justifying usage of machine learning methods for missense mutations effect explanation. In the end their effectiveness is compared to the output of knowledge & physical potential methods. Predictions are made on tertiary structure basis, out of which in case of a few MetaMisTher components FASTA amino acid sequence is extracted.

Our meta-predictor provides $\Delta(\Delta G)$ values by wrapping 7 local physical potential, knowledge-based potential and machine learning methods into a common independent server framework. To increase precision of $\Delta(\Delta G)$ scoring, MetaMisther uses information from.PDB files and predicts solvent accessibility and secondary structure. The gathered predictions are processed with a consensus SVM algorithm to obtain a single score and evaluated with ROC curves and MCC (Matthew's correlation coefficient).

References:

1. Khan S (2010) *Mutational effects on protein structures: Knowledge gained from databases, predictions and protein models*; University of Tampere, PhD thesis.
2. Olatubosun A *et al* (2012) *Human Mutation* **33**: 1166–1174.

Posters

P4.1

An efficient approach for estimating GMM initial conditions as a way of improvement of Nuclear Magnetic resonance spectra analysis

Franciszek Binczyk[1], Michał Marczyk[1], Andrzej Polanski[2], Joanna Polanska[1]

[1]Silesian University of Technology in Gliwice, Data Mining Group, Gliwice, Poland; [2]Silesian University of Technology in Gliwice, Institute of Informatics, Gliwice, Poland

e-mail: Franek Binczyk <franciszek.e.binczyk@polsl.pl>

Nuclear magnetic resonance (NMR) spectroscopy is a popular technique used for estimation of chemical compounds amount in a complex mixture. The analysed signal has a form of spectrum with peaks representing different compounds. Peaks are commonly not well separated. To identify them there is a need for robust technique of signal decomposition; promising approach is to use Gaussian mixture model (GMM). Parameters of GMM components, like mean value, variance and weight, can be estimated with use of Expectation-Maximization algorithm. The crucial problems in EM are: finding initial conditions and number of model components. Aim of the study was to construct efficient technique for initial conditions estimation in Expectation-Maximization (EM) algorithm to improve analysis of NMR spectroscopy.

To solve this problem authors propose to apply technique that was already successfully used for decomposition of proteomic spectra. It is based on division of spectrum into intervals dependant on defined threshold for peak heights. In each interval EM is used to decompose a part of the signal limited by interval borders. The maximal number of GMM components for this partial decomposition is not strictly defined but is greater or equal than the number of peaks detected in given interval with use of local maxima method. Optimal number of components is found with use of Bayesian information criterion (BIC). The procedure is repeated for each interval and in consequence results of partial decompositions are used as initial condition for global spectrum analysis.

The efficiency of algorithm was examined on a 27 data sets collected with human brain phantom. Examined spectrum consists signals emitted by chemical compounds, like NAA, Choline, Creatine, Glutamine, My-inositol, Phospho-creatine, Lactate and other. In this study we compared proposed approach with an analysis done on whole spectrum at once. For all datasets differences in estimated amount of all compounds compared to true values were observed. The smallest increase of accuracy: 2–5% (average for 27 spectra equals 3.91%) was observed for compounds that are represented by well-separated peaks such as NAA. For close multiplets such as lactate or glutamine, or highly overlapped peaks such as myo-inositol and choline the increase of accuracy was more than 10% (average for 27 spectras equals 14.5%).

In this preliminary study we proved that approach for estimation of EM initial conditions, that was successfully used in MALDI-TOF spectra analysis, may be applied, however with minor adjustments which are specific to characteristics of NMR spectra.

Acknowledgements

Authors are thankful to the group of Medical Physics of Prof Maria Sokol at Centre of Oncology in Gliwice for access to the data that were used in this work.

This work was financed by internal grant for young researchers from Silesian University of Technology in Gliwice (BKM RAu 2014 t.16).

P4.2

Multiobjective optimization of mutation accumulation in bacterial genomes

Paweł Błażej, Paweł Mackiewicz, Małgorzata Grabińska

Faculty of Biotechnology, University of Wrocław, Wrocław, Poland
e-mail: Pawel Blazej <blazej@smorfland.uni.wroc.pl>

The process of mutation accumulation is an important component of biological evolution. Together with selection, it is responsible for genetic variation of organisms and their adaptation to changing environments. Therefore, mutations should be desired. On the other hand, it seems that most of mutations introduced into genomes are deleterious. Then, we should expect a decrease of mutation rate to minimize this harmful effect. As a result of this, some kind of trade-off should be observed between the necessity to accurately replicate the established genetic information and the requirements for adaptational flexibility of organisms. To answer the question about the optimality of the mutation accumulation process, we carried out Monte Carlo computer simulations applying evolutionary strategy and multiobjective optimization. Using this approach we tried to find such nucleotide substitution process that would be optimal for assumed criteria. The mutation introduction was modelled by a Markov process described by the General Time Reversible model assuming a fixed stationary distribution and codon usage as in several bacterial genomes. As optimization criteria we examined: probability of nucleotide substitution, cost of amino acid substitution and probability of stop codon occurrence. Based on these assumptions we found a class of mutation substitution matrices that produced the smallest number of harmful mutations and simultaneously generated sufficient genetic variation. The simulations showed that there are many optimal solutions representing a trade-off between the considered objectives. The results were compared with the empirical mutational matrices which were found in real genomes.

P4.3

Bioinformatics analysis of rat *Mbl1*gene reveals enrichment of transcription factor binding sites distinctive in promoters of acute-phase proteins genes

V. Bondarenko[1,2], M. Yu Obolenska[1]

[1]Institute of Molecular Biology and Genetics NAS, Laboratory of Systems biology, Ukraine; [2]ESC "Institute of Biology", Taras Shevchenko National University, Kiev, Ukraine
e-mail: Vladyslav Bondarenko <vlad-sage@yandex.ua>

Background: The mannose-binding lectin (Mbl1) is a protein of complement cascade and an inherent part of the innate immune system. The complement cascade represents a complex network of plasma proteins that cooperate to withdraw the pathogens and the "old" cells from the circulation and to maintain healthy tissue. This network integrates both the antibody- and cell-driven responses to infection by pathogens. Whenever MBL is bound to specific molecular patterns on the specified cells it attracts MBL-associated serine proteases (MASPs) and *via* them initiates complement pathway activation (Schwaeble *et al*., 2011) with eventual opsonisation of particles, release of inflammatory peptides, enhanced engulfment by phagocytes of above-mentioned cells as well as cell debris and thrombin-like activity (Takahashi *et al*., 2011). However, a little is known about it's transcriptional regulation. To get more in-depth knowledge on regulation of *Mbl1* transcription we have undertaken the *in silico* search for transcription factors binding sites (TFBS) in promoter of this gene in *Rattus norvegicus.*
Methods: The promoter length was 1500 bp (from –1500 to + 73 bp), containing adjacent 5'UTR of *Mbl1* genes of *Rattus norvegicus* (Gene ID: 24548) and *Mus musculus* (Gene ID: 17194). The position - weight matrices of transcription factor binding sites were obtained from the database Matrix Family Library Version 9.0. The program MatInspector was used for the search. The phylogenetic analysis was conducted by the comparison of the promoter of *Rattus norvegicus* with the promoter of gene-ortologue of *Mus musculus*. The functional modules of presumable transcription factors were determined with ModelInspector program.
Results: As a result of Mbl1 promoter analysis we obtained 14 liver-specific and evolutionary conserved TFBS of *Mbl1* gene of *Rattus norvegicus* located predominantly within the region of 1000 bps in length. This list includes such transcription factors as ISGF3 (Interferon-stimulated growth factor), members of PARbZIP (proline and acidic amino acid-rich basic leucine zipper) transcription factors family, FKHD (forkhead box proteins), HNF1 and HNF6 (hepatocyte nuclear factors), glucocorticoid receptors (GR) and CCAAT/enhancer binding proteins (C/EBP), AP-1 (activator protein 1) and ETS, E26 transformation-specific transcription factor. Additionally, we analyzed found transcription factors for possible functional interactions by searching for modules in ModelInspector program. Both single transcription factors and modules are typical for the promoters of the genes encoding the proteins of acute phase response, such as fibrinogen, factor VIII and factor IX (Begbie, 1999), other genes of complement system. Thus, as the result of this work we found similar patterns of conservative cis-elements in promoter of Mbl1 gene. This analysis hints bisical assumtions about possible mechanism of Mbl1 gene expression and differenciate candidates amond TFs for further task-oriented experimental validation.

P4.4

The mitochondrial outer membrane protein import complexes of Amoebozoa representatives

Dorota Buczek[1,2], Małgorzata Wojtkowska[1], Yutaka Suzuki[3], Seiji Sonobe[4], Yukinori Nishigami[4,5], Hanna Kmita[1], Wojciech Makałowski[2]

[1]Laboratory of Bioenergetics, Adam Mickiewicz University, Poznań, Poland; [2]Institute of Bioinformatics, University of Muenster, Muenster, Germany; [3]Department of Medical Genome Sciences, The University of Tokyo, Kashiwa, Japan; [4]Department of Life Science, Graduate School of Life Science, University of Hyogo, Harima Science Park City, Hyogo, Japan; [5]Department of Physics, Kyoto University, Japan
e-mail: Dorota Buczek <wojmak@uni-muenster.de>

The TOB/SAM and TOM complexes are translocases located in the outer membrane of mitochondria and responsible for protein import into mitochondria. It is well known that the complexes contain core subunits formed by proteins common to all eukaryotes and additional subunits that appeared later in the evolution of eukaryotes and are lineage-specific. In the Amoebozoa the subunit organization of the TOM and TOB/SAM complexes is not very well understood. Here, we present putative organization of these complexes in several amoebozoans: *Acanthamoeba castellanii, Amoeba proteus, Dictyostelium discoideum, Dictyostelium purpureum, Dictyostelium fasciculatum and Polysphondylium pallidum*. We used publicly available genome and transcriptome data and in two cases (*A. castelanii* and *A. proteus*) mRNA sequences determined by us.
The assembly of transcritpomes was performed using the Trinity assembly suite. Then, all the genome and transcriptome sequences were scanned for the TOM and TOB/SAM complexes coding sequences. In this step the BLAST search algorithm was employed with several references from different taxonomical lineages retrieved from the GenBank. In some cases Hidden Markov Model based searches were employed as well.
The amoebozoan TOM complex appears to luck Tom5 and Tom6 as we couldn't detect the encoding sequences in any amoebozoan genome. However, we cannot exclude that their homologs exist in these organisms and their detection was precluded by their size and applied methodology. All the other TOM complex subunits were predicted, though not all of them in all the genomes. In the case of the TOB/SAM complex most of the subunits were detected across the genomes of species analyzed with the exception of two entamoebas whose TOB/SAM complex seems to consist of Sam50 proteins. In general, entamoebas have both complexes slimmed down, which may reflect their adaptation to parasitic/commensal life style. Taking together, the amoebozoan TOM and TOB/SAM complexes contain core subunits but also display some diversity of subunit organization that suggests the organization changes occurring during their evolutionary history.

Acknowledgements

The studies were supported by the project performed within the "MPD Programme" of Foundation for Polish Science, cofinanced by European Union, Regional Development Fund (Innovative Economy Operational Programme 2007-2013), MPD/2010/3.

P4.5

Multi-parameter sensitivity analysis based on the information theoretic measure

Agata Charzyńska[1], Michał Komorowski[2], Anna Gambin[3]

[1]Institute of Computer Science, Polish Academy of Sciences, Poland; [2]Division of Modelling in Biology and Medicine Institute of Fundamental Technological Research, Polish Academy of Sciences, Poland; [3]Institute of Informatics, University of Warsaw, Warsaw, Poland
e-mail: Agata Charzynska <a.charzynska@phd.ipipan.waw.pl>

The modelling of biochemical reaction networks is an important element of modern systems biology that aims to describe and explain various natural phenomena. All theoretical models which describe any complex system consists of many parameters and many variables depending on the parameters. To utilize a model it is necessary to determine the parameters values, not only for model setting but also for understanding of various interactions within the modelled species. The proper sensitivity analysis, that enables investigation of the impact of parameters to the model response, is crucial to model validation. Therefore it is critical to introduce an accurate computational method for the SA. Recently available methods for multiparameter model analysis are being developed but still infancy.
We introduce the method based on an information theoretical measure called mutual information. The method hierarchically determines the subgroups of parameters that have the crucial significance to the model. To quantify the sensitivity of the model, we randomly search the parameter space, seeking the parameters with biggest weights based on sensitivity indexing. To our knowledge, no sensitivity analysis of interactions within groups of many parameters can be performed computationally efficiently, limiting our ability to compare the influence of possibly correlated parameters. This new approach makes significant improvements to available methods by not only including information about the dynamics of the model depending on single parameters but also the answer for the specific questions like, what is the common influence of the parameters to the model response, what are the thresholds of the parameters values for the different system behavior. Moreover, this simple method enables to capture interactions within a model regardless of the existence of correlations between parameters.

P4.6

RetrogeneDB — a database of animal retrogenes

Michał Kabza, Joanna Ciomborowska, Izabela Makałowska

Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznan, Poznań, Poland
e-mail: Joanna Ciomborowska <joannac@amu.edu.pl>

Retrogenes are copies of existing genes that arise from insertion of reverse transcribed mRNAs into the genome. Because of the lack of regulatory elements, which are not inherited, such copies used to be considered as non-functional (retropseudogenes) and classified as "junk DNA". Multiple discoveries, most notably in the past 10 years, have challenged that view. Since then, multiple mechanisms by which retrogenes may become functional have been proposed. A unique feature of retrogenes – the loss of all or most of the cis-regulatory elements, leads to the lower level of evolutionary constraint, which is the reason why retrogenes may relatively easily undergo neofunctionalization. As a result, retroposition is a vital part of the process of development of lineage- and species-specific traits.
Although multiple attempts have been made to detect retrogenes in the genomes of model organisms, there is still no repository of retrogenes for a broader range of organisms. Here, we present a new database called RetrogeneDB (http://retrogenedb.amu.edu.pl/) that contains high-quality retrogene datasets for 62 genomes from Ensembl release 73. The search was done based on the similarities between reference genomic sequence and proteins coded by multiexon genes in a given species. To increase accuracy, we applied several criteria to call a genomic region a retrocopy: length of the alignment at least 150 bp, minimum of 50% coverage of parental gene, minimum of 50% identity, and loss of at least two introns among others.
Our strategy led to identification of 84,808 retrocopies, including 6,277 protein-coding genes not recognized previously as retrogenes. A total of 64,225 retrocopies identified by us are not present in the Ensembl database, this includes 139 retrocopies in the human and as many as 2,205 in the mouse genome, which belong to the best annotated. Because of our stringent requirements, applied in the order to generate a high-quality data set, the number of identified retrocopies in a given species is considerably lower than in most other databases. However, this method gave consistently good results in both, well and poorly annotated, low-coverage genomes, for example, alpaca or dolphin.
RetrogeneDB allows users to search for retrogenes and their parental genes using numerous criteria, such as genomic localization, key words, parental gene name, and retrocopy ID. Results can be filtered based on the retrocopy type, open reading frame conservation or expression which was estimated for selected organisms based on RNA-Seq data. In addition, a JBrowse genome browser was implemented allowing retrocopy inspection in the genomic context. The search from parental gene perspective enables to identify all retrocopies of a given gene or all orthologs, which were retroposed in any other species. Users can also perform sequence-based search using BLAST tool.

Acknowledgments

This work was supported by Polish National Science Center under projects DEC-2013/09/N/NZ2/01221.

P4.7

SimRNA: employing long range contact related intra-chain entropy and experimentally calibrated base pairing potentials in 3D RNA structure prediction

Wayne K. Dawson[1], Michal Boniecki[1], Janusz M. Bujnicki[1,2]

[1]International Institute of Molecular and Cell Biology, Warsaw, Poland; [2]Institute of Molecular Biology and Biotechnology, Poznań, Poland
e-mail: Wayne Dawson <wdawson@genesilico.pl>

In RNA and protein structure prediction, typically threading and homology approaches are used. However, to succeed, threading approaches require known structures (or structures with sufficient similarity). Moreover, threading does not provide any insight on the dynamic properties of the predicted structures and does not explain the changes in stability of RNA structures that result from mutations and evolution of RNA. This can only be comprehended by understanding the core of the thermodynamics.
SimRNA is a recently developed de novo 3D structure prediction program in our laboratory that uses the Monte Carlo method to search the conformation space of RNA using knowledge based energy functions. To proceed with the development of thermodynamic potentials, the study has been focusing on the following objectives. First, we have been tuning the SimRNA knowledge-based energies to approach the free energy of the experimentally obtained Turner energy rules obtained from short RNA duplexes. These measured parameters provide an independent standard on which to calibrate the SimRNA force field. Second, we have introduced a physical model to re-rank the energies of SimRNA based on the entropy effects of intra-chain "contacts" (or "cross links") resulting from RNA base pairing interactions. The cross linking entropy method is similar to the contact order model but is a *systematically quantitative* calculation approach — not merely qualitative. It has been shown to be far more accurate in predicting the stability of evolved RNA sequences: in particular, familiar known structures such as tRNA and rRNA 5S.
The methodology often yields improvements comparable to the first level of clustering and is a genuine physical model. It is expected that it can be applied universally to both de novo 3D RNA structure prediction, 3D protein structure prediction, and particularly, to predicting the structural stability in evolution and mutation experiments.

P4.8

"In silico" prediction of tRNA and Trm10 methyltransferase complexes

Michał Dyzma, Irina Tuszyńska, Janusz M. Bujnicki

International Institute of Molecular and cell Biology in Warsaw, Laboratory of Bioinformatics and Protein Engineering, Warsaw, Poland
e-mail: Michał Dyzma <mdyzma@genesilico.pl>

Transfer RNA (tRNA) methylation is necessary for the proper biological function of tRNA. During tRNA maturation, extensive specific post-transcriptional modifications are introduced by many different enzymes to ensure proper tRNA structure and function. One of the most common modifications is the m^1N9 modification, which is catalyzed by Trm10, S-Adenosyl-L-homocysteine dependent methyltransferase. Trm10 enzyme group include several homolougues, which transfer methyl group to: guanine(9)-N(1), adenine(9)-N(1) and guanine(9). TRM10 homologues are widely found in eukaryotes and archaea, but not in bacteria. Here we present computational analysis of *Sulfolobus acidocaldarius* initiator tRNA and Trm10-transferase monomer docking complexes obtained *"in silico"*.
Using method for 3D homology modeling, developed in Bujnicki's lab (modeRNA), we have obtained model of tertiary structure of studied tRNA sequence. Sequence was optimized using Quick Refinement of Nucleic Acids tool developed in the same lab. To generate possible orientations and conformations of the components, we employed the GRAMM method and produced 10000 decoys for each experimental group. Obtained sets of decoys were scored and grouped using FILTREST3D according to distance restrains between specific residues of the protein and tRNA (SAH-N9; K-64-tRNA; Y161-tRNA; K254-tRNA). Residues were chosen due to their conserved character and electric potential properties. Cut-off function allowed to pick out best structures, which were submitted to further analysis with DARS-RNP allowing to identify structures close to the native.

P4.9

Fast and accurate *de-novo* DNA genome assembly algorithm

Wojciech Frohmberg[1,2], Michał Kierzynka[1,3], Paweł Wojciechowski[1,3], Piotr Żurkowski[1], Jacek Błażewicz[1,3]

[1]Institute of Computer Science, Poznan University of Technology, Poznań, Poland; [2]Institute of Plant Genetics, Polish Academy of Sciences, Poland; [3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland
e-mail: Wojciech Frohmberg <wojciech.frohmberg@cs.put.poznan.pl>

Next generation sequencers are more and more widely spread all over the world facilitating sequencing genomes of organisms that could not have been sequenced before, but also re-sequencing already known organisms to examine heterogeneity of a single species. In fact, because of its accessibility the number of application for sequencing technology is rapidly growing. The increased use of sequencers leads inevitably to overload the servers that produces genomes in a process of DNA assembly. The solution of the problem could be the acceleration of the DNA assembly methods. On the other hand assembly still needs to be as accurate as possible to be able to answer biomedical questions and verify scientific hypothesis.
To satisfy mutually opposite criteria of high accuracy and equally high performance we have implemented new DNA assembly method. The method utilizes the overlap-layout-consensus strategy in order to ensure high quality results and new technologies, in particular, GPU acceleration to produce results efficiently. The talk will focus on the algorithmic side of the method with particular emphasis on the genome graph traverse step.

P4.10

Distributed and parallel reconstruction of gene regulatory networks with BNFinder

Alina Frolova[1], Bartek Wilczynski[2]

[1]Institute of Molecular Biology and Genetics of NASU, Systems Biology Laboratory, Ukraine; [2]Institute of Informatics, University of Warsaw, Computational and Systems Biology Group, Warsaw, Poland
e-mail: Alina Frolova <fshodan@gmail.com>

Bayesian networks are probabilistic graphical models widely used for gene regulatory networks reconstruction, because they are able to infer non-linear relationships between genes. In general, learning Bayesian networks from experimental data is NP-hard, leading to widespread use of heuristic search methods giving suboptimal results. However, it was showed by Dojer [1] that it is possible to find optimal network in polynomial time when datasets are finite and there are external constraints ensuring network acyclicity. While our method makes it possible to reconstruct optimal networks in polynomial time [2], the widespread adoption of the algorithm is limited by its long running times bound by the time it takes to find the optimal set of parents for the most complex variable.

In this work we present a new and improved version of BNFinder — our tool for learning optimal Bayesian networks. The improvement consist of parallelized inference algorithm providing significant speedup with good efficiency. Importantly, the improved implementation is quite insensitive to heterogeneous datasets, i.e. situations where complex variables with multiple parents are mixed with simple variables. We show algorithms performance measured on simulated datasets as well as real biological data regarding phosphorylation network inference [3]. Finally, we adapted our tool to heterogeneous distributed environment and performed genome-scale *E. coli* network inference [4] using Ukrainian GRID infrastructure.

References:

1. Dojer N (2006) Learning Bayesian networks does not have to be NP-hard. *Math Found Comput Sci* 305–314.
2. Wilczynski B, Dojer N (2009) BNFinder: exact and efficient method for learning Bayesian networks. *Bioinformatics* **25**: 286–287.
3. Sachs K, Perez O, Peer D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**: 523–529.
4. Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biol* **5**.

P4.11

Simulation experiments for shotgun DNA coverage processes

Mateusz Garbulowski[1], Andrzej Polański[2]

[1]Institute of Automatic Control, Silesian University of Technology, Katowice, Poland; [2]Institute of Informatics, Silesian University of Technology, Katowice, Poland
e-mail: Mateusz Garbulowski <mateuszgarbulowski@gmail.com>

The purpose of the study was creating a computer simulation system for the shotgun sequencing processes. The main aims related to elaborated system was to [1] obtain simulation results for better understanding and verifying the shotgun sequencing process [2] analysis of the influence of some effects ignored in the Lander-Waterman theory on the statistics of the shotgun sequencing process [3], analyses of several real sequencing datasets.

Main methods which we applied to create the model, were based on the realistic shotgun sequencing methods such us randomly creating reads (the short nucleotide fragments) or adaptation the model to the parameters such as read length (L), the length of whole nucleotide fragment (G) and the numbers of reads (N). The final step of algorithm was counted number of contigs (full nucleotide sequences, that was created by placing the reads) and their length, that was used to analysis parameters with comparing it to the Lander-Waterman statistics [2].

To further explore the shotgun sequencing method we collected the real data, which contains many sequences of reads coming from the *1000 Genomes Project*. These data was collected in .BAM files and we took the reads and put together to create a contigs sequences, according to the information of position to the reference sequences of each read. To make this we use *Matlab* environment and the main model modification. We plotted the number of contigs results as the relationship between σ parameter [3] (the ratio of empirical number of contigs to the maximum predicted value of contigs) to the "a" parameter (which means depth of coverage understood as NL/G). The created graph was compared to the Lander-Waterman statistic.

The next step was making a plot which show another relationship of parameter σ to the tandem repeating sequences which was created in shotgun sequencing process. This data which we used to create the plot was searched on the NCBI database. The main information, which we needed was placed in section *Assembly Statistics* (for each chromosome) and in the attached articles. We collected the data for each chromosomes of chosen organisms, that we put together on one graph.

The above analysis let to study the shotgun sequencing process in so many ways. Created model was able to tell us the basic parameters influence to the final result. Obtained plot show the visible relationship of σ parameter and depth of coverage „a", which show the crucial downward trend of σ parameter in order to rising parameter „a". The chart, that present relationship between σ and number of tandem repeats show us the dependence of those two parameters. Given results lead to better knowledge of shotgun sequencing method, which inform how to manage a better quality of this method and study the probability processes of contig formation.

References:

1. Polański A, Kimmel M (2007) *Bioinformatics*. Springer 243–252.
2. Lander ES, Waterman MS (1988) *Genomics* **2**: 231–239.
3. Wendl MC, Yang S-P (2004) *Bionformatics* **20**: 1527–1534.

P4.12

Searching for mutational matrices in bacterial genomes

Małgorzata Grabińska, Paweł Błażej, Paweł Mackiewicz

Department of Genomics, Faculty of Biotechnology, University of Wrocław, Wrocław, Poland
e-mail: Małgorzata Grabińska <ewan@smorfland.uni.wroc.pl>

The mutational pressure is one of the most important machinery that propels evolutionary process in each organism. Therefore, the knowledge about mutation patterns acting in genomes is necessary for understanding evolution and variability of organisms. The aim of our study was finding mutational nucleotides substitution matrices in many different bacterial genomes. The nucleotide substitutions were calculated separately for orthologs lying on differently replicated DNA strands (leading and lagging) from closely related bacteria belonging to the same species or genus. In order to eliminate the effect of selection and to obtain the purest as possible mutational pressure, we eliminated highly expressed protein genes based on Codon Adaptation Index (CAI) parameter minimalizing type I and type II errors. Only fourfold degenerated positions in codons were used to the final calculation of substitutions matrices to further avoid an influence of selection pressure. The matrices were calculated using both maximum parsimony (Fitch's algorithm) and maximum likelihood methods. The obtained matrices differed across the studied genome sets although the same nucleotide substitutions characterized usually by big rate. Stationary distributions resulting from the Markov process described by the matrices were not always in equilibrium with the nucleotide composition of the studied sequences in every genome set.

P4.13

Apple miRNAs and their role in Fire Blight resistance

Elzbieta Kaja[1], Timothy McNellis[2], Michal Szczesniak[1], Izabela Makalowska[1]

[1]Laboratory of Bioinformatics, Faculty of Biology, Adam Mickiewicz University Poznań, Poznań, Poland; [2]Department of Plant Pathology, Penn State University, PA, USA
e-mail: Elzbieta Kaja <eo@amu.edu.pl>

Micro RNAs (miRNAs) are small, single stranded RNA molecules, which are involved in post-transcriptional gene silencing in plant and animal cells. To date, it has been reported that plant miRNAs, by targeting many regulatory genes, play an important role in such processes as: plant development, hormone signaling or biotic and abiotic stress response. Although, many interesting facts have already been discovered about miRNA nature and way of action, those molecules are still surprising and not fully understood.

In this research we characterized miRNAs, which are specific for Gala apple scions grafted on four different rootstocks: B.9, G.30, M.27 and M.111, presenting diverse Fire Blight resistance. Our previous studies (Jensen *et al.*, 2009) showed, that those rootstocks also induce a different gene expression pattern in the apple scion as well as they determine tree size. Although the mechanism of this regulation is not known yet, we suggest that miRNAs might play a crucial role in it.

In order to identify miRNA species, as well as their expression levels in selected trees, SOLiD sequencing of small RNAs has been performed. All the reads have been mapped to the apple genome (http://www.rosaceae.org/projects/apple_genome) and searched for conserved and apple-specific miRNAs. Performed analyses allowed us to extend the apple miRNA repertoire by 38 conserved and 78 novel, apple specific, miRNA as well as verify 143 miRNAs from previous studies. We confirmed five of new miRNAs using qPCR or RT-PCR. We also identified miRNAs with significantly changed expression among analyzed rootstocks. Rootstock-dependent fold change in miR expression levels, together with qPCR confirmations let us define five apple miRNAs potentially involved in fire blight resistance in apple trees: miR169a, miR319c, miR160e, miR167b-g and miR168a,b. In addition, we searched for potential miRNA targets using psRNATarget focusing on transcripts with significantly higher expression in fire blight resistant trees.

P4.14

The impact of the time delay between heat shock treatment and chemo- or radiotherapy on suppression of NF-κB pathway response

Małgorzata Kardyńska, Jarosław Śmieja

Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
e-mail: Małgorzata Kardyńska <aijoku@gmail.com>

Multimodal oncological strategies have been recently the subject of increasing interest. This type of therapy may consist in combining chemotherapy or radiotherapy with hyperthermia and is proposed to improve the efficacy of anticancer treatments. However, a good understanding of the cell regulatory mechanisms is necessary to develop proper therapy protocols.

Both of these therapeutic procedures, heat shock and chemo- or radiotherapy are associated with activation of number of proteins and genes through particular signaling pathways, among which HSF and NF-κB pathways seem to be the most important. Activation of the HSF pathway occurs as a result of heat shock and leads to production of HSPs (Heat Shock Proteins), known to prevent protein misfolding. On the other hand, chemo- or radiotherapy induce NF-κB activation which regulate the transcription of hundreds of genes, including genes that determine cell fate, and can play an antiapoptotic role in cancer cells. It is also known that after a heat shock the response of the NF-κB pathway is damped for some time, which may explain the increased sensitivity of cells to chemo- or radiotherapy after heat shock treatment.

In our work we propose a combined model of those two pathways, and we use it to determine the time delay after which chemo- or radiotherapy should begin, to get the best possible results. The model was developed on the basis of existing ones, but in order to add crosstalk between the HSP and NF-κB pathways, they had to be modified. In the proposed model two types of excitations may be applied: heat shock and TNF stimulation (which corresponds to irradiation). The interactions between the HSF and NF-κB pathways take into account creation HSP:IKK complexes. It has been reported that HSP72 may associate with the IKKγ/NEMO subunit of the IKK complex, which lead to inhibition of IKK activity and seems to be most important mechanism for suppression the NF-κB pathway response.

By carrying out a number of simulation we were able to determine the time window in which the most effective suppression of the NF-κB pathway response is observed. For the TNF stimulation initiated between 4 and 6 hours after start of heat shock treatment, the NF-κB response is most efficiently dumped. This effect corresponds to the highest level of free HSPs in the cell's cytoplasm, which appears after 4.5 hours from the beginning of heat shock. The simulation results are consistent with the results of experimental studies gathered by our group.

Acknowledgements

This work was partially supported by NCN grant DEC-2012/05/B/NZ2/01618.

P4.15

Secondary structure based algorithm for manganese (II) coordination spheres prediction on 3D structures of proteins

Tatyana A. Khrustaleva[1], Vladislav V. Khrustalev[2], Eugene V. Barkovsky[2]

[1]Institute of Physiology of the National Academy of Sciences of Belarus, Laboratory of Cellular Technologies, Belarus; [2]Belarussian State Medical University, Department of General Chemistry, Belarus
e-mail: Tatyana Khrustaleva <tanissia.lir@gmail.com>

The algorithm entitled "VVTAK Mn(II)" (http://chemres.bsmu.by) is based on four propensity scales created during the study on 3D structures of 149 bacterial proteins [1] and 300 proteins of chordates with coordinated Mn (II) cations. The first scale is based on the specific secondary structure elements distribution around Asp, Glu and His residues coordinating Mn (II) ions. Indeed, as we found out, 77.88% of all the coordination spheres from bacterial proteins contain at least one residue in the specific "beta strand – binder – random coil" motif [1]. Both beta strand and alpha helix may be situated after that motif. The second scale is based on amino acid propensities in five positions before and five positions after the binding residue. The third scale is made from propensities for pentapeptides composed of hydrophobic and hydrophilic amino acids to be situated near the residue interacting with Mn (II). The fourth scale is based on frequencies of amino acid combinations in pairs around the binding residue. The forth scale helps to exclude from predictions those potential binders which are surrounded by positively charged residues.

The algorithm performs the search for Asp, Glu and His residues situated around each of the "active binders" predicted by a combination of probability scales in the three dimensional space. All those amino acids situated near the "active binder" are included into the coordination sphere (or in the "ion trap"). Recommended distance for that search is equal to 6 Angstroms. Recommended minimal number of amino acids required to form a coordination sphere is equal to 3. We used 30 proteins of plants, archaea, fungi and invertebrates as training set. At the recommended parameters sensitivity and specificity of the algorithm are both equal to 74.3%.

The algorithm is written as an extended Ms Excel spreadsheet. To use the algorithm one has to enter the entire text from PDB file and the amino acid sequence in FASTA format as an input. The output is the text that should be saved as PDB file. To visualize atoms participating in Mn (II) ions binding one may choose options "Temperature" (in "Display" menu) and "Ball & Stick" (in "Colours" menu) in the RasMol viewer. Moreover, the algorithm provides a table with predicted "active binders", as well as a table with predicted coordination spheres.

Reference:

Khrustaleva TA (2014) *Advances in Bioinformatics* **2014**: 1–14.

P4.16

Spatial model of stress-induced transposon proliferation

Anna Gambin[1], Dariusz Grzebelus[2], Mateusz Kitlas[3], Arnaud Le Rouzic[4], Michał Startek[5]

[1]Institute of Informatics, University of Warsaw, Warsaw, Poland; [2]Faculty of Horticulture, University of Agriculture in Krakow, Kraków, Poland; [3]Institute of Informatics, University of Warsaw, Warsaw, Poland; [4]Laboratoire Évolution, Génomes et Spéciation, Centre National de la Recherche Scientifique, France; [5]Institute of Informatics, University of Warsaw, Warsaw, Poland
e-mail: Mateusz Kitlas <mateusz.kitlas@gmail.com>

Evolution is not a continuous process, it is composed of long stabilization periods and intensive changes which result in adaptation to the new niche and speciation. During the stabilization periods, most activity which causes structural changes in the genome negatively affects the survivability of specimens and of the whole population.
However, in presence of environmental stress caused for example by climate change, or by colonization of new environmental niches, introduction of new sources of variability may be beneficial. Mobile genetic elements (transposons) are DNA segments capable of changing their genomic localization. Such activity is potentially dangerous for host organism, therefore during the stability periods, there are several mechanisms controlling the transposition, such as RNA interference, DNA methylation, histone modification.
Here, as in previous paper, we try to show, that the appearance of stress results in suppression or elimination of these mechanisms, what enables the explosion of transposition activity resulting in structural changes in genomes.
Such changes are undirected and increase, in average, the distance from the current phenotypic optimum. However, in the situation of changing environment they could promote the beneficial genetic innovation on the scale of whole population.
Here, we present a stochastic computational model including spatial effects of transposon proliferation in asexual populations. The model is a spatial extension of work previously presented in the paper "Genomic parasites or symbionts? Modeling the effects of environmental pressure on transposon activity in asexual populations" (Theoretical Population Biology). The model uses a Fisher geometric phenotypic landscape with standard Gaussian selection. Mutations are Gaussian as well, but they are affected by transposon activity in order to enable us to study, the interplay between transposition rate and environmental stress. Several scenarios are studied, among them a geographical spread of population to new environments, a colonization of new subniche by an already established species and evolutions of conditions of gradually changing environment.
The strength of this model lies in the fact that, unlike in previous transposition-selection equilibrium-based models, there is no direct relation between the transposon count and the organism's fitness function (the only link is the transposons' influence on the mutation rate).
We shall present how the environmental stress and spatial effects stimulate transposon activity.
In particular, we will present how our model predicts several real-world phenomena, such as increased transposon counts in species colonizing new environments as well as increased transposon counts on the frontline of colonization wave.

P4.17

Computational modeling of amyloid peptides

Malgorzata Kotulska

Wroclaw University of Technology, Institute of Biomedical Engineering and Instrumentation, Wrocław, Poland
e-mail: Malgorzata Kotulska <malgorzata.kotulska@pwr.edu.pl>

Amyloids are proteins capable of forming fibrils whose intramolecular contact sites form a characteristic zipper pattern. A number of diseases related to amyloid proteins is constantly increasing and include Alzheimer's disease (proteins Abeta and tau), Parkinson's disease (alpha-synuclein), type 2 diabetes (amylin), Creutzfeldt-Jakob's disease (prion), Huntington's disease (huntington), amyotrophic lateral sclerosis (SOD1), and many others. Recognition of factors responsible for protein misfolding and subsequent cascade of events can contribute to better understanding of the diseases mechanisms and potential drug design. Recent studies indicate that short segments of aminoacids, which are believed to be 4-10 residues long and called emph{hotspots}, can underly amyloidogenic properties of a protein. Those fragments can be harmless when they are burried inside a protein. Other studies indicate that the neurodegenerative processes is related only to transient amyloid oligomers and may correspond to their incorporation into cell membranes, creating weakly cation-selective ion channels that allow uncontrolled influx of calcium or other ions into nerve cells. In the talk we will present our results of computational methods that recognize amyloidogenic hotspots and their propensity to forming double strands, based on classical machine learning methods, probabilistic formal grammars, and our original method based on site specific co-occurrence pattern (Stanislawski, 2013, Kotulska, 2013, Gasior, 2013). Preliminary results on modeling hypothetical structures of the amyloid pores will also be discussed.

References:

Stanislawski J, Kotulska M, Unold O (2013) *BMC Bioinformatics* **14**: 21.
Kotulska M, Unold O (2013) *BMC Bioinformatics* **14**: 351.
Gasior P, Kotulska M (2014) *BMC Bioinformatics* **15**: 54.

P4.18

Development of sparse matrix crystal screens dedicated to RNAs and protein-RNA complexes crystallization

Lukasz P. Kozlowski[1], Radoslaw Pluta[1], Astha[1], Janusz M. Bujnicki[1,2]

[1]International Institute of Molecular and Cell Biology, Laboratory of Bioinformatics and Protein Engineering, Poland; [2]Institute of Molecular Biology and Biotechnology, Faculty of Biology, Laboratory of Bioinformatics, Poland
e-mail: Lukasz Kozlowski <lukaskoz@genesilico.pl>

Conditions leading to the crystallization of a RNA and protein-RNA complexes can be generally identified by screening a broad range of chemical mixtures. To narrow down the search, the data mining of the RNA structures deposited in the Protein Data Bank (PDB) [1] can be utilized. To date, only protein [2], protein-DNA [3], and nucleic acid [4] sparse matrix crystallization screens were commercially available.
As of May 2014, PDB contained 2843 structures with RNA molecules, among which 1032 were solely RNA structures. To remove bias in this dataset, all ribosomal structures were deleted. Similarly, all redundant structures (99% and more sequence identity) were removed using CD-HIT [2]. This protocol led to 462 structures from which data representing crystallization conditions (pH, temperature, buffers, precipitants e.g. salts, PEGs, additives e.g. spermine) were extracted, clustered and analyzed.
According to the data, the majority of RNA-protein complexes crystallize at range pH 6–7, while RNA alone prefers pH ~7 (slightly acidic to prevent RNA hydrolysis). Additionally, the presence of NaCl, magnesium salts, and precipitants such as PEG 4000 or ammonium sulfate dominated the chemical space. In respect of temperature, most RNA structures were obtained at room temperature or 37°C, while for RNA-protein complexes this restriction was weaker and many of them crystallized at temperature below 20°C. Those observations allow us to the design of the first sparse matrix crystallization screen dedicated to the protein-RNA complexes and to the formulation an up-to-date screen for RNAs crystallization.

References:

1. Berman J *et al* (2000) *Nucleic Acids Res* **28**: 235–242.
2. Smialowski P *et al* (2010) *Methods Mol Biol* **609**: 385–400.
3. Pryor E *et al* (2012) *Acta Crystallogr Sect F Struct Biol Cryst Commun* **68**: 985–993.
4. Kondo J *et al* (2014) in *Handbook of RNA Biochemistry*, Wiley, chapter 23.
5. Li W *et al* (2001) *Bioinformatics* **17**: 282–283.

P4.19

Molecular characteristics recognition of estrogen receptor beta selective agonists

Paweł Książek, Krzysztof Bryl

University of Warmia and Mazury in Olsztyn, Chair of Physics and Biophysics, Olsztyn, Poland
e-mail: Paweł Książek <pawel.ksiazek@uwm.edu.pl>

Estrogen receptor β (ERβ), a member of nuclear receptor superfamily, is a molecule with therapeutic potential for several human diseases (Shanle & Xu, 2010). Its ligand-binding cavity is very similar to that present in ERα. Furthermore, the receptor cavities of both ERα and ERβ are relatively flexible, and depending on the nature of the bound ligand, the shape of the cavity may change significantly (Koehler *et al*., 2005). This makes ERβ a challenging target for new potent drugs. Despite the difficulties, considerable advances have recently been made and new synthetic ERβ agonists exist (Minutolo *et al*., 2011). However still little is known about structural mechanisms behind this successes. In order to reveal selectivity mechanisms and to design new potent ERβ agonists computational methods can be utilized (Spyrakis & Cozzini, 2009).
We carried out a molecular docking study of selected ERβ agonists. The ligands were docked to carefully chosen ERβ ligand binding domain structures obtained from the RCSB Protein Data Bank (PDB). This study showed how structural features of agonists such as alkyl substituent length, its aliphatic chain disubstitution and molecule stereochemistry can influence subtype selectivity. This knowledge will support the design of new ligands exhibiting sufficient levels of ERβ selectivity for therapeutic use.

References:

Koehler *et al* (2005) *Endocr Rev* **26**: 465–478.
Minutolo *et al* (2011) *Med Res Rev* **31**: 364–442.
Shanle, Xu (2010) *Adv Drug Deliv Rev* **62**: 1265–1276.
Spyrakis, Cozzini (2009) *Current Medicinal Chemistry* **16**: 2987–3027.

P4.20

Initial comparative genomics analysis of bionanocellulose producing strain: *Gluconacetobacter xylinus* E25

Katarzyna Kubiak[1], Aleksandra Kot[1], Marta Kurzawa[1], Marzena Jędrzejczak-Krzepkowska[1], Mariusz Krawczyk[2,3], Andrzej Migdalski[2], Magdalena M. Kacprzak[2], Damian Loska[2,3], Karolina Ludwicka[1], Marek Kołodziejczyk[1], Przemysław Rytczak[1], Alina Krystynowicz[1], Stanisław Bielecki[1]

[1]Institute of Technical Biochemistry, Lodz University of Technology, Łódź, Poland; [2]Genomed S.A., Warsaw, Poland; [3]Polish-Japanese Institute of Information Technology, Warsaw, Poland
e-mail: Katarzyna Kubiak <katarzyna.kubiak@p.lodz.pl>

Complete genome sequence of the first cellulose synthesizing *Ga. xylinus* strain was announced recently (Kubiak *et al*., 2014). It was assembled by the means of Next Generation Sequencing technology by Genomed S.A. The species is used worldwide in industrial production of bionanocellulose (BNC)-based cosmetic masks and wound dressings (Czaja et al., 2006). The most intensive research is done in the field of development of scaffolds based on BNC to be used in various tissues regeneration, with the most advances in cartilage and bone replacements (Gama *et al*., 2012; Kowalska-Ludwicka *et al*., 2013).

The genome of *Ga. xylinus* E25 consists of one 3.4 Mbp chromosome and five plasmids (from 2.2 to 336.1 kb in size). We found it quite interesting to identify a megaplasmid (pGX5), which was not reported for any *Gluconacetobacter* representative up to date. However, megaplasmids were quite commonly reported in other *Alphaproteobacteria* genera. In these known cases they usually carry genes essential for cell survival in unfavorable environment. Some details about pGX5 encoded genes' annotation will be presented.

Our results are of biotechnological importance due to the industrial application of the sequenced species but as only two other *Gluconacetobacter* representatives genomes sequences are known up to date (namely: *Ga. xylinus* NBRC 3288 strain, defective in cellulose production (Ogino *et al*., 2011) and *Ga. hanseni* ATCC 23769 strain (Iyer *et al*., 2010)) they bring important scientific novelty as well. Therefore initial comparative analysis were performed with these genomes' sequences.

Initial automated annotation performed with Prodigal, DIYA and PIPA resulted in putative function assignment to 72.5% of all predicted chromosome and plasmid protein-coding genes. Further analysis, including metabolism modelling (with Pathologic package) allowed for better insight in pathways active in the species. Initial results including new GGDDEF/EAL family of proteins members (essential for c-di-GMP signaling) will be presented.

References:

Czaja W *et al* (2006) *Biomaterials* **27**: 145–151.
Iyer *et al* (2010) *J Bacteriol* **192**: 4256–4257.
Gama M *et al* (2012) *Perspectives in Nanotechnology* **9:** 157–174. Eds. Boca Raton, Florida: CRC Press.
Kowalska-Ludwicka *et al* (2013) *Arch Med Sci* **9:** 527–534.
Kubiak K *et al* (2014) *J Biotech* **176:** 18–19.
Ogino *et al* (2011) *J Bacteriol* **193:** 6997–6998.

P4.21

Retrocopies — they are not „junk" DNA any more

Joanna Ciomborowska, Magdalena Kubiak, Michał Kabza, Izabela Makałowska

Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznan, Poznań, Poland
e-mail: Magdalena Kubiak <mrakubiak@gmail.com>

Retrocopies are RNA-based duplicates originated from reverse transcription of mRNA and incorporation of cDNA into a genomic sequence. This process is called retroposition and it results in a formation of a single-exon copy from a multi-exon parental gene. Majority of retrocopies are usually inactive and therefore are commonly called retropseudogenes or just pseudogenes. For many years they have been considered as useless, so called "junk DNA". Nevertheless, since the discovery of the first functional retrogene in 1985 the interest in retrogenes (expressed retrocopies) increased and subsequent studies revealed that this type of sequences is important from the evolutionary point of view. Now we have more and more examples showing retrocopies as a driving force in evolution of animals and playing an important role in shaping interspecies differences. Also our studies reveled that we should reconsider retropseudogenes and look at them from a new perspective as we still underestimate the real number of animal retrogenes.

We decided to check whether previously annotated as pseudogenes retrocopies can be transcribed and be in fact functional. Candidates for our analyses come from developed by us database of animal retrogenes — RetrogeneDB (retrogenedb.amu.edu.pl). Taking into account two main criteria, such as estimated expression based on RNA-Seq analysis and status in our database we choose 68 human and 16 mouse retrocopies for expression verification utilizing PCR and pooled human and mouse cDNA libaries as templates.

We were able to confirm expression of 41 human and 11 mouse retrogenes. For those retrogenes we performed a set of bioinformatics analyses in order to pinpoint their putative functions. Among them we identified 29 human and 9 mouse genes with disrupted open reading frames, which together with confirmed expression, strongly suggests neofunctionalization. Moreover significant part of examined retrogenes is specific for one organism. 32% of expressed human and 90% of expressed mouse retrogenes has no orthologs in other organisms. This observation shows how big could be the impact of retroposition on variation among species.

P4.22

3D Modeling of Group I Intron structures by comparative modeling with ModeRNA and *de novo* RNA folding with SimRNA

Kumar Deepak[1], M. J. Boniecki[2], Janusz M. Bujnicki[2,1]

[1]Laboratory of Structural Bioinformatics, Institute of Molecular Biology and Biotechnology, Collegium Biologicum, Adam Mickiewicz University, Poznań, Poland; [2]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland
e-mail: Deepak Kumar <deepak@genesilico.pl>

Group I introns is a family of widespread non-coding RNA molecules well known for self-splicing from the host precursor RNA. Thus far only Azoarcus, Tetrahymena and Twort are known and well studied structures of this family. Commonly, group I introns are classified into 14 subfamilies [1] based on conserved core sequences and peripheral structures. However, introns from particular groups have high length diversity and weak sequence similarity, which makes structure prediction for these RNAs very difficult.
To provide 3D structural models of representatives of group I introns from all families, we used a combination of comparative and de novo RNA structure modeling. We manually prepared alignment of 11 representatives (from subfamilies with unknown structures) with sequences and structures of representatives with known structures. ModeRNA [2] software was used to generate initial models of group I intron core structures by a comparative modeling approach. We defined the structural core based on the available secondary and tertiary structures, P4-P6 domain containing P4, P5 and P6 and P3-P9 domain containing P3, P7, P8 and P9. Azoarcus, Tetrahymena and Twort sharing significant similarity and common secondary structure with the representatives were chosen as templates for modeling. Fragments of models without counterparts in templates were then added and folded with a *de novo* modeling approach, as implemented in the SimRNA method (Boniecki, Bujnicki, and coworkers, manuscript in preparation).
The generated models of group I intron structures accurately depict the global topology, secondary and tertiary interactions. Expectedly, the accuracy is highest in the core, with RMSD between 3–4 Å, whereas deviations are larger for peripheral regions that differ substantially between different introns. The results of this analysis provide a 3D perspective for studying group I introns and for interpretation of their sequence evolution in a structural context.

References:

1. Zhou Y, Lu C, Wu QJ, Wang Y, Sun ZT, Deng JC, Zhang Y (2007) GISSD: Group I Intron Sequence and Structure Database. *Nucleic Acid Res* 2008; 36 (Database issue): D31–D37.
2. Rother M, Rother K, Puton T, Bujnicki JM (2011) RNA tertiary structure prediction with ModeRNA. *Briefing in Bioinformatics* (2011) **12**: 601–613.

P4.23

Mathematical model of bystander effect

Karolina Kurasz, Krzysztof Łakomiec, Aleksandra Krzywoń, Krzysztof Fujarewicz

Silesian University of Technology, Institute of Automatic Control, Gliwice, Poland
e-mail: Karolina Kurasz <karolina.kurasz@polsl.pl>

Aim: The aim of this study is to propose a mathematical model of bystander effect and results of parameter estimation based on biological experiments. The model describes ways in which cell survival is influenced by the rate of neighbouring cells. There are two different effects described in literature. The first is the classical Bystander effect, where cell survival is reduced when communicating with irradiated cells. The second, recently reported, is reciprocal bystander effect where we observe increase in cell survival of irradiated cells close to unirradiated bystander cells.
Materials and methods: Ultraviolet radiation is considered to be one of the most important etiological factors of skin cancer. On the basis of the mechanisms of action, it appears that the bystander effect induced by UV radiation can have a share of damaging (carcinogenesis) and possibly skin cells in other tissues. The Bystander effect of ultraviolet (UV) radiation is the biological phenomenon in which unirradiated cells exhibit radiation effects as a result of molecular signals received from nearby cells irradiated. Ultraviolet radiation generates reactive oxygen and nitrogen species from irradiated cells which induce bystander effects in unirradiated cells, such as reduction in clonogenic cell survival and delayed cell death and apoptosis. The current investigations are based on biological experiments and computer simulations to make clear the biology of bystander effect. The in vitro experiments were performed using human dermal fibroblasts exposed to UV. Irradiated and unirradiated cells were co-incubated in six-well dishes with an insert separating two population of cells by a 0.4-mm pore membrane to allow diffusion of medium components and molecular signals between them [1].
Results: The model specifies the dynamics of two cells populations and is described by five ordinary differential equations. The model assumes that normal cells (N) are characterized by unlimited population growth. Subpopulation of damaged and non proliferating cells (D) appears after irradiation and due to native DNA breaks. Molecular signals transmitted *via* factors (β) released into the medium are generated by irradiated cells. They are inhibited by normal cells. β factors influence the rate of normal cells damage and the repair rate. Numerical estimation of the parameters of mathematical model is not a trivial task, because there is no universal method of performing this process. Furthermore, standard numerical methods of searching minimum of the objective functions are strongly dependent on the starting point. Therefore, the process must be repeated multiple times using different initial valuesof the parameters. This considerably increases the time of calculations. The parameters of the model were initially estimated using ADFIT program- a tool for numerical parameter estimation [2]. For further parameters estimation we used sensitivity analysis. As a result we obtained the model, which is well fitted to the experimental data. Moreover, the model is able to explain both classical and reciprocal effects.

References:

1. Widel M, Krzywon A, Gajda K, Skonieczna M, Rzeszowska-Wolny J (2014) Induction of bystander effects by UVA, UVB, and UVC radiation in human fibroblasts and the implication of reactive oxygen species. *Free Radic Biol Med* **68**: 278–287.
2. Łakomiec K, Fujarewicz K (2014) Parameter estimation of non-linear models using adjoint sensitivity analysis. *Advanced Approaches to Intelligent Information and Database Systems* 59–68.

Acknowledgement

The work was financially supported by NCN grant register number DEC2012/05/B/ST6/03472.

P4.24

Energy terms of the mirror images in protein structure prediction

Monika Kurczyńska, Ewa Kania, Bogumił Konopka, Małgorzata Kotulska

Institute of Biomedical Engineering and Instrumentation, Wroclaw University of Technology, Wrocław, Poland
e-mail: Monika Kurczyńska <monika.kurczynska@pwr.edu.pl>

Recently, more and more research about protein structure prediction are based on the residue-residue contact information such as contact maps [1, 2]. Tools for protein structure reconstruction from contact map generate the model collection containing the properly oriented models and their mirror images, because all of them share the same contact map. The properly oriented models and their mirror images can constitute competitive forms in the nature, therefore may have the same total energy. In our work we investigated 100 models of the 311 domains from class A and 183 domains from class B from SCOP database [3] reconstructed with C2S_pipeline [2]. We calculated the structural features of the models and their energy terms with PyRosetta [4].

The numbers of mirror images and properly oriented models in a collection were similar. One of the most often used indicators for protein structure quality is the Root Mean Square Deviation (RMSD). However, we observed that not always high RMSD indicated mirror images. We found a lot of domains whose properly oriented and image model had similar RMSD values. Despite this we observed that some of the structural features could be useful in separating correctly oriented and mirror models. One of them was the torsion angle φ, which allowed distinguishing native and mirror models for more than 60% of protein. Despite the fact that the total energy was not always different for both type of models, we showed that some of the energy terms (e.g.: regarding torsion angles, attractive and repulsive forces, solvation process) could be used to qualify the model with regard to its orientation.

References:

1. Duarte JM *et al* (2010) *BMC Bioinformatics* **11**: 283.
2. Konopka BM *et al* (2014) *J Membr Biol* **247**: 409–420.
3. Murzin AG *et al* (1995) *J Mol Biol* **247**: 536–540.
4. Chaudhury S *et al* (2010) *Bioinformatics* **26**: 689–691.

P4.25

Differentiation of ampicillin-resistant and ampicillin-sensitive *Escherichia coli* strains using infrared spectroscopy and multilayer perceptrons

Łukasz Lechowicz[1], Mariusz Urbaniak[2], Wioletta Adamus-Białek[3], Wiesław Kaca[1]

[1]Institute of Biology, Department of Microbiology, The Jan Kochanowski University, Kielce, Poland; [2]Institute of Chemistry, Organic Chemistry Division, The Jan Kochanowski University, Kielce, Poland; [3]Independent Department of Environmental Protection and Modeling, Jan Kochanowski University, Kielce, Poland
e-mail: Łukasz Lechowicz <lechowiczlukasz@gmail.com>

Introduction: In recent years, it can be seen an increased interest in the use of infrared spectroscopy and artificial neural networks (e.g. multilayer perceptrons) in the differentiation of bacterial strains to the species and subspecies level. The aim of this work was to use the multilayer perceptrons and the spectroscopic data (bacterial infrared spectra) for the differentiation of ampicillin-resistant and ampicillin-sensitive uropathogenic *Escherichia coli* strains. **Materials and Methods**: In this work 127 *E. coli* strains were used. The ampicillin susceptibility tests were performed according to the EUCAST standards. The bacteria were cultured on LB agar at 37°C for spectroscopic analysis. After 24 hours of incubation for each strain 20 typical bacterial colonies were collected. The colonies were placed on crystal of ATR/FT-IR spectroscope. Each sample was scanned 25 times and the results were averaged. In the next step a number of multilayer perceptrons were designed for differentiation of the bacterial strains based of their infrared spectra. **Results:** From all tested *E. coli* strains 73 (57.5%) were sensitive to ampicillin and 54 (42,5%) were resistant to the tested antibiotic. The best multilayer perceptrons differentiate *E. coli* strains with an accuracy of 90%. **Conclusions**: Infrared spectroscopy and artificial neural networks allow to identify of the antibiotic resistance of bacterial strains.

Acknowledgments

This project was financed from the funds of the National Science Center allocated on the basis of the decision number DEC-2012/07/N/NZ7/01187.

P4.26

MODOMICS: a database of RNA modification pathways—2013 update

Magdalena A. Machnicka[1], Kaja Milanowska[1,2], Okan Osman Oglou[3], Elzbieta Purta[1], Malgorzata Kurkowska[1], Anna Olchowik[1], Witold Januszewski[1], Sebastian Kalinowski[2], Stanislaw Dunin-Horkawicz[1], Kristian M. Rother[1,2], Mark Helm[3], Janusz M. Bujnicki[1,2], Henri Grosjean[4]

[1]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland; [2]Faculty of Biology, Adam Mickiewicz University, Poznań, Poland, [3]Institut für Pharmazie und Biochemie, Johannes Gutenberg-Universität, Mainz, Germany; [4]Centre de Génétique Moléculaire, UPR 3404, CNRS, Université Paris-Sud, FRC 3115, Gif-sur-Yvette, France
e-mail: Magdalena Machnicka <mmika@genesilico.pl>

MODOMICS is a database of RNA modifications that provides comprehensive information concerning the chemical structures of modified ribonucleosides, their biosynthetic pathways, RNA-modifying enzymes and location of modified residues in RNA sequences. In the current database version, accessible at http://modomics.genesilico.pl, we included new features: a census of human and yeast snoRNAs involved in RNA-guided RNA modification, a new section covering the 5'-end capping process, and a catalogue of 'building blocks' for chemical synthesis of a large variety of modified nucleosides. The MODOMICS collections of RNA modifications, RNA-modifying enzymes and modified RNAs have been also updated. A number of newly identified modified ribonucleosides and more than one hundred functionally and structurally characterized proteins from various organisms have been added. In the RNA sequences section, snRNAs and snoRNAs with experimentally mapped modified nucleosides have been added and the current collection of rRNA and tRNA sequences has been substantially enlarged. To facilitate literature searches, each record in MODOMICS has been cross-referenced to other databases and to selected key publications. New options for database searching and querying have been implemented, including a BLAST search of protein sequences and a PARALIGN search of the collected nucleic acid sequences.

P4.27

Carotenoids intercalation and arrangement in phosphatydylcholine bilayer – a molecular dynamics simulation study

Krzysztof Makuch, Michał Markiewicz, Marta Pasenkiewicz-Gierula

Jagiellonian University, Department of Computational Biophysics, and Bioinformatics, Kraków, Poland
e-mail: Krzysztof Makuch <krzysztof.makuch@uj.edu.pl>

Carotenoids are produced in plants and some other photosynthetic organisms [1]. As animals are incapable of synthesizing carotenoids, they must obtain them through their diet. Carotenoids perform various functions in prokaryotic and eukaryotic organisms. In animals, they play a protective role against oxidative stress, probably have an anticancer activity, and are a major group of vitamin A. In humans, among the most abundant carotenoids are lutein and zeaxanthin, which can be found in high concentration in the fovea of the eye retina [2].

As was shown in experimental studies [3], both lutein and zeaxanthin exist in liposomes not only in monomeric form, but also as aggregates. Such aggregates, due to the limited mobility, cannot reorient as freely as monomers.

The primary aim of this study is verification of the arrangement of lutein and zeaxanthin in the 1-palmitoyl-2-oleoyl-phosphocholine (POPC) bilayer using molecular dynamics (MD) simulation method. Our earlier MD simulations [4] confirmed experimental results that lutein can be both in perpendicular and parallel orientation relative to the lipid bilayer surface [4]. That study was limited to a relatively small number of intercalated lutein molecules in the POPC bilayer. In this study, a larger number of systems was analyzed. This study shows that lutein intercalates into the bilayer with both its ends (β and ε rings) and its initial orientation is either perpendicular or parallel to the normal; initial orientations of zeaxantin are similar.

Vital role during carotenoids insertion plays formation of hydrogen bonds between the carotenoid hydroxyl group and the POPC phosphate group.

In this study, both intercalation of aggregates into the POPC bilayer and their formation already in the bilayer are observed.

Acknowledgements

All systems were simulated with GROMACS molecular dynamics package, using OPLS-AA forcefield. This research was supported in part by PL-Grid Infrastructure.

Reference

1. Dall'Osto L, Lico C, Alric J, Giuliano G, Havaux M, Bassi R (2006) *BMC Plant Biol* **6**.
2. Richer S, Stiles W, Statkute L, Pulido J, Frankowski J, Rudy D, Pei K, Tsipursky M, Nyland J (2004) *J Am Optom Assoc* **75**: 216–229.
3. Sujak, Oukulski, W, Gruszecki WI (2000) *Biochim Biophys Acta* **1509**: 255–263.
4. Pasenkiewicz-Gierula M, Baczyński K, Murzyn K, Markiewicz M (2012) *Acta Biochim Pol* **59**: 115–118.

P4.28

Modeling of MALDI-ToF spectra by parallel computing

Michal Marczyk[1], Joanna Polanska[1], Andrzej Polanski[2]

[1]Data Mining Group, Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland; [2]Institute of Informatics, Silesian University of Technology, Gliwice, Poland
e-mail: Michał Marczyk <Michal.Marczyk@polsl.pl>

Introduction: MALDI-ToF mass spectrometry allows characterization of low-molecular-weight fractions of the human proteome in a relatively short time, thus it emerges as a valuable tool during disease treatment. Major fragments of the mass spectrum are signal peaks, which correspond to the proteins contained in the analyzed sample. In our study mass spectrum is decomposed into a sum of Gaussian bell-shaped curves by using a variant of the expectation maximization (EM) algorithm. Introduced algorithm allows to use fast graphics processor units (GPUs) or multicore central processor units (CPUs) to increase computational speed.

Results: We tested efficiency of different implementations of EM algorithm using real data and simulated spectra with different length and number of peaks. In all cases using simple precision calculations speeds-up GPU implementation. Computational time of algorithm increases with a complexity of the spectrum, indicating a greater advantage of parallel computing. For datasets with high number of peaks we can get more than tenfold speed boost using Tesla graphics card comparing to single CPU. We get similar increase for datasets with high number of points per spectrum.

Conclusions: Parallel computing approach enables efficient and intuitive implementation of algorithm for decomposition of mass spectrum. Parallelization of the code drastically speed-ups modeling algorithm with comparison to the standard implementation on a single CPU. Obtained results depend on the type of hardware used. The most efficient implementation was based on mixed strategy, maximizing usage of available CPUs and GPUs.

Acknowledgments

This work was funded by internal grant of Silesian University of Technology for young researchers BKM 2014.

P4.29

Heterogeneity of cellular immune response for TNF-α stimulation

Joanna Markiewicz, Bogdan Kaźmierczak, Marek Kochańczyk, Tomasz Lipniacki

Institute of Fundamental Technological Research of Polish Academy of Sciences, Deparment of Mechanics and Physics of Fluids, Warsaw, Poland
e-mail: Joanna Markiewicz <jciesiel@ippt.pan.pl>

Heterogeneity of immune response is observed among cell lines for different stimulants. There are a lot of sources of cellular variability. Intrinsic noise is associated with stochastic gene expression influencing fluctuations of RNA and protein levels. Extrinsic noise is caused by differences in cellular state, cell cycle or micro-environment. In our experiment we focus on NFκB signaling pathway and its activation *via* TNF-α (tumor necrosis factor) stimulation.

Reference:

Satija R, Shalek A.K. (2014) *Trends in Immunology* Vol. 35, No. 5.

P4.30

RNApathwaysDB — a database of RNA maturation and decay

Kaja Milanowska[1], Katarzyna Handing[3], Anna Lukasik[3], Marcin Skorupski[3], Zuzanna Hojszyk[3], Magdalena A. Machnicka[2], Martyna Nowacka[2], Kristian M. Rother[3,2], Janusz M. Bujnicki [2,3]

[1]Department of Gene Expression, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland; [2]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland; [3]Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland
e-mail: Kaja Milanowska <kaja@amu.edu.pl>

Many RNA molecules undergo complex maturation, involving e.g. excision from the primary transcripts, posttranscriptional modification, splicing, and polyadenylation. The level of mature RNAs in the cell is controlled by degradation, which proceeds *via* many different reactions including, but not limited to endo- and exonucleolytic cleavage. The systematization of data about RNA metabolic pathways and enzymes taking part in RNA maturation and degradation is essential for the full understanding these processes. RNApathwaysDB (RNA pathways database) is the first database of metabolic pathways involving RNA as the substrate. It presents information about reactions and enzymes (proteins, RNA molecules or complexes) that take part in RNA processing. The database provides also links to other databases and literature information. The current dataset is limited to the maturation and degradation of tRNA, rRNA and mRNA, and describes pathways in three model organisms: *Escherichia coli*, *Saccharomyces cerevisiae* and *Homo sapiens*. Other RNAs, enzymes and pathways and data for other organisms will be successively added in the future. The database can be queried with keywords or by the name of a pathway, a reaction, an enzymatic complex, a protein or an RNA molecule. Amino acid sequences of protein enzymes involved in pathways included in RNApathwaysDB can be compared to a user-defined query sequence with a BLAST utility. Options for data presentation include graphs of pathways and tabular forms with enzymes and literature data. Structures of macromolecular complexes are presented as "potato models" using DrawBioPath – a new javascript tool. The contents of RNApathwaysDB can be accessed through the World Wide Web at http://genesilico.pl/rnapathwaysdb.

P4.31

EvOligo — an oligonucleotide grouping software for designing DNA/RNA hybridization-based high-throughput assays

Marek C. Milewski[1], Karol Kamel[1], Anna Kurzyńska-Kokorniak[1] Marcin K. Chmielewski[1], Marek Figlerowicz[1,2]

[1]Institute of Bioorganic Chemistry Polish Academy of Sciences, Poznań, Poland; [2]Institute of Computing Science, Poznań University of Technology, Poznań, Poland
e-mail: Marek Milewski <akurzyns@man.poznan.pl>

Experimental methods based on DNA/RNA hybridization such as Multiplex polymerase chain reaction (Multiplex PCR) or Multiplex ligation-dependent probe amplification (MLPA), require the use of mixtures of multiple oligonucleotide molecules, primers or probes, in a single test tube. To provide an optimal reaction environment, one must achieve minimal self- and cross-hybridization between those oligonucleotides. To meet this problem we developed EvOligo, a software package for grouping and designing short RNA and DNA sequences. EvOligo allows for a construction of groups of molecules, characterized by the weakest possible cross- or self-interactions, by sorting critical parts of the sequences into sets, and designing the rest of the sequences with the use of evolutionary algorithm. Grouping can also be made based on melting temperature of duplexes. EvOligo uses a nearest-neighbor model of nucleic acids interactions and a parallel evolutionary algorithm that allow the software to take an advantage of modern multi-core CPUs. Applications of the groups and sequences generated with the software include but are not limited to designing MLPA or Multiplex PCR experiments as well as constructing special-purpose microarrays.

Acknowledgements

This work was partially supported by the European Union Regional Development Fund within the Innovative Economy Programm [Grant No. POIG.01.03.01-30-045/09].

P4.32

Bioweb: framework for applications used for genetic data analysis

Robert Nowak

Institute of Electronic Systems, Warsaw University of Technology, Warsaw, Poland
e-mail: Robert Nowak <rbmnowak@gmail.com>

A characteristic feature of the computer programs applied to genetic data is the necessity to analyze large amounts of data using complex algorithms, which means thathigh performance is crucial. Different user and system requirements mean that the flexibility of software is also important. Finally, users prefer a graphical interface that is accessible from a web browser and applications that update automatically.

Scientists are becoming increasingly involved in software development. They should use software engineering practices and tools to avoid common mistakes and to speed up the development tasks. The architecture of working application with explanation of development decisions could help in developing new computer programs.The use of appropriate systems also facilitates rapid prototyping, which allows us to verify concepts by obtaining the requisite information from end users: biologists and doctors.

In this study, we describe the BioWeb framework i.e. application architecture, the programming languages, libraries and tools,used to develop applications for processing genetic data. High performance, flexibility, and a user interface with a web browser are achieved by using multiple programming languages.The software used by our framework and the framework itself were created with C++, Python, and JavaScript with HTML5.Our solution is similar to frameworks that connect C++ with Python or Python with JavaScript, but BioWeb combines three programming languages.

A lot of bioinformatics frameworks and libraries exists for C++ (NCBI C++ Toolkit), Java (BioJava), Python (BioPython), R (Bioconductor) and other programming languages. All these solutions impose limitations connected with the usage of only one programming languageand do not support the user interface in a web browser. The heavyweight web-based genome analysis frameworks, such as Galaxy, have a lot of ready-made modules and meet most of the requirements for systems for the genetic data analysis. Our framework allows you to create smaller and independent solutions, which are easier to manage end to customize. It could be easily extended to use GPU and/or computing clusters, which is required in production-scale analysis.

The presented framework was used to build a number of genetic data processing applications. It is available from the website http://bioweb.sourceforge.netunder LGPLv3. The project website includes 'getting started' instructions, user manual and example applications.

P4.33

Statistical methods for integrating high-throughput biological data

Anna Papiez[1], Chrstophe Badie[2], Joanna Polanska[1]

[1]Institute of Automatic Control, Silesian University of Technology, Katowice, Poland; [2]Centre for Radiation, Chemical and Environmental Hazards, Public Health England, United Kingdom
e-mail: Anna Papież <anna.papiez@polsl.pl>

The development of biological high-throughput measurement techniques created the possibility of obtaining large amounts of information in a single experiment. However, these methods often are costly, require a lot of time and provoke problems typical for high-dimensional data analysis. These issues imply the need for effective procedures for integrating existing data sets stored in bioinformatics databases and laboratories.
This study consisted of testing the impact of integrating two independent microarrray expression data sets using p value combination statistical techniques. The method was proposed as an alternative for biological and functional validation of a single gene expression set in the case when we have more than one set of data available and the design of both experiments is compatible.
The data was obtained from two microarray experiments designed for the search of biomarkers of radiosensitivity in breast cancer patients. These studies were consistent in terms of design, but were conducted of different microarray platforms. After the preprocessing stage involving normalization and batch effect removal, the p-values resulting from multiple testing procedures applied in order to find differentiating genes were combined. This was done using a variety of transformations in order to find the one that is most suitable for the analyzed data sets. The tested procedures included Fisher's, Lancaster's and modifications of Stouffer's methods for the combination of p-values.
The weighted Z-score algorithm (one of the modifications of Stouffer's method) using the inverse standard error as weight proved to be the most appropriate method for integrating the investigated data sets using p-value combination in terms of the nature of the data as well as biological conclusions drawn. The analysis shows that p-value combination is an efficient method of integrating independent data sets. However, choice of the used transformation should strictly rely upon the knowledge of analyzed data.

Acknowledgements

The work was financially supported by NCN grant HARMONIA 4 register number 2013/08/M/ST6/00924 (JP, CB) and SUT - BKM/524/RAU1/2014.t17. Additionally, AP is holder of scholarship DoktoRIS – Scholarship program for Innovative Silesia. Calculations were carried out using GeCONiI infrastructure (POIG.02.03.01-24-099/13).

P4.34

Prediction of coagulation factor XI dimeric structure in mammals

Michal B. Ponczek

University of Lodz, Faculty of Biology and Environmental Protection, Department of General Biochemistry, Łódź, Poland
e-mail: Michał Ponczek <mponczek@biol.uni.lodz.pl>

The coagulation factor XI (FXI) is a novel protein in vertebrate evolution, present only in mammals offering an extra way for more effective blood clotting. The consequence of an increased risk of thrombosis is a grim side effect. FXI takes part in intrinsic blood coagulation after factor XII (FXII) activation on negatively charged contact surfaces. Additionally, the activation of FXI is caused by thrombin and a positive feedback boosts the generation of additional thrombin. FXI is a consequence of prekallikrein (PK) duplication, a protein which is present in all land vertebrates. The two proteins are structurally alike as both have four apple domains (A1, A2, A3, A4) and one serine protease domain (S1). The variation is that FXI forms homodimers which adapted to be an effective factor IX (FIX) activator. Human FXI structure was resolved by X-ray crystallography (PDB: 2F83) and a biological assembly of the dimer was also calculated. It revealed how the homodimer is formed on molecular level by bunch of amino acids forming interfaces stabilized by hydrogen bonds, hydrophobic interactions and one covalent bond between two Cys321. X-ray crystallography of the whole PK is not resolved but from the sequence and spatial predictions it is known that its failure to form dimers results from differences of the region corresponding to the interface. Furthermore, establishing the interchain S-S covalent bond is not possible as an intrachain bond is formed between Cys321 and Cys326. Contrarily, FXI has glycine in 326 position which enables interchain bond formation between two Cys321. FXI 2F83 can be used as template to model structures of homologs from other mammals. The purpose was to predict FXI dimeric structure of mammals for better understanding of dimer interface interactions developed through evolution for the function of this protein. Selected sequences of mammals were used for automated homology modeling in Swiss-Model (http://swissmodel.expasy.org/) to build dimeric spatial structures of hypothetical proteins on 2F83. A structure of mammalian PK dimer was also modeled on 2F83 to stress differences between FXI and PK. The interface fragments of the models were conformationally optimized in molecular modeling software (Scigress V3.1.6). The top of the PK dimer interface misses hydrogen bonds and the bottom lacks hydrophobic and electrostatic interactions. Respectively, predicted FXI mammalian dimers resemble the 2F83 biological assembly in interfacial interactions. The interesting case are rabbit and platypus dimers. Both lacks Cys321 mutated into other amino acids but the rest of their dimer interface is properly stabilized by electrostatic and hydrophobic interactions. The S-S interchain bond seems to be a stabilizer of the dimer but is not necessary for the dimerization of FXI which is enabled by evolutionary original electrostatic and hydrophobic interfacial interactions.

P4.35

RNA-Seq data analysis pipeline in Poznan Supercomputing And Networking Center

Juliusz Pukacki, Hubert Świerczyński, Cezary Mazurek, Michał Kosiedowski

Poznan Supercomputing and Networking Center, Poznań, Poland
e-mail: Juliusz Pukacki <pukacki@man.poznan.pl>

High-Throughput DNA Sequencing, which is typically referred to as RNA-Seq, is becoming the most important tool for gene expression analysis. Besides of specialized equipment for performing sequencing on biological material it also involves computing infrastructures for data storage and data processing.
Poznan Supercomputing and Networking Center (PSNC) is an institution aiming to provide networking, computational and storage environment to support scientific communities, locally in Wielkoposka region area, and globally by participating in Polish and European initiatives. PSNC is also very active in the research projects area. One of the examples of that kind of activity is support for Greater Poland Cancer Centre in the RNA-Seq data analysis.
For performing RNA-Seq analysis, resources that are part of Pl-Grid infrastructure were utilized. Pl-Grid is national hardware and software platform grouping together resources located across Poland, design and implemented to enable research in various domains of e-Science. The platform (QosCosGrid) offers advanced job and resource management capabilities to deliver to end-users supercomputer performance. Taking advantage of resource management interfaces, in collaboration with MD Anderson Cancer Center, automatic pipeline for RNA-Seq analysis was designed and implemented.
It consist of five main steps:
— Alignment: short sequences (reads) are aligned to the reference genome
— Compression: intermediate step responsible for compressing SAM file to BAM format
— BAM Transformation: processing of BAM file to prepare input for further computations (grouping, reordering, sorting, duplicates detection)
— Reads Counting: in this step reads are calculated and stored
— Quality Control: computation of a series of quality control metrics
Simultaneously with standard RNA-Seq workflow the additional research was done on retroelements analysis. For this purposes, preparation of dedicated unmasked reference genome was required.
The results of the workflow was then used for further statistical processing, mostly focused on differential expression assays.
Initially, the designed computational pipeline was applied for data coming from several biological experiments, including reprogramming human somatic cells towards induced pluripotent stem cells (iPS cells) and analyzing the role of TRIM28 knockdown in this process. Another set of data was generated from analysis of TRIM28 role in breast cancer homeostasis.
Conducted experiments proved, that properly designed computational pipeline, integrated with storage capabilities, can significantly facilitate work with RNA-Seq data. The infrastructure provided by PSNC is very scalable and can easily fulfill requirements of biomedical communities. The other important aspect is ability to build interdisciplinary teams, for solving scientific problems coming from different research areas.

P4.36

Construction of biomolecular computer with DNA and restrictions enzymes

Joanna Sarnik[1], Sebastian Sakowski[2], Janusz Błasiak[1], Tadeusz Krasiński[2], Tomasz Popławski[1]

[1]University of Lodz, Department of Molecular Genetics, Łódź, Poland; [2]University of Lodz, Department of Algebraic Geometry and Theoretical Informatics, Łódź, Poland
e-mail: Joanna Sarnik <jsarnik@biol.uni.lodz.pl>

Biomolecular computers built of nucleic acids, along with quantum computers, would be an alternative for traditional, silicon based computers. Main advantages of biomolecular computers are massive parallel processing of data, expanded capacity of storing information and compatibility with living organisms. However, biomolecular computers have several drawbacks including time-consuming procedures of preparing of input, problems in detecting output signals and interference with by-products. Due to this obstacles, there are few laboratory implementations of theoretically designed DNA computers, but there are many implementations of DNA computers for particular problems.
We have constructed a biomolecular computer with double stranded DNA molecules which implements a simple model of computer – the finite automaton. All elements of a finite automata: input words, states and transitions are made of DNA molecules. Restriction enzymes together with DNA ligase act as hardware of this computer. All the DNA molecules were joined in one reaction tube. Alternating cutting of input words and ligating of transition molecules (in a programmed fashion) reflect the action of finite automata. The computer stop working by creating a long DNA molecule representing a final state. It is detected by the electrophoresis. This construct is an enhanced multistate version of the Shapiro automaton. We have optimized our biomolecular computer by establishing optimal reaction conditions and successfully testing three different input words. This approach may be used (in the future) to build nanomachines (made of DNA molecules) which may be applied in medicine, pharmacy or biotechnology.

Acknowledgements

This project is supported by the National Science Centre of Poland (NCN). Decision: DEC-2011/01/B/NZ2/03022.

P4.37

The interaction between matrix metalloproteinases and their tissue inhibitors in normal and disease conditions

Beata Sokołowska[1], Krystiana A. Krzyśko[1,2], Irena M. Niebroj-Dobosz[1], Marta Hallay-Suszek[2], Łukasz Charzewski[2], Agnieszka Madej-Pilarczyk[1], Michał Marchel[3], Irena Hausmanowa-Petrusewicz[1], Bogdan Lesyng[1,2]

[1]Mossakowski Medical Research Center Polish Academy of Sciences, Warsaw, Poland; [2]University of Warsaw, Warsaw, Poland; [3]Medical University of Warsaw, Warsaw, Poland
e-mail: Beata Sokolowska <beta.sokolowska@imdik.pan.pl>

Background: Metalloproteinases (MMPs) and their tissue inhibitors (TIMPs) are supposed to take part in some neurological disorders and may have diagnostic and prognostic values. In particular, altered levels of MMPs and their TIMPs are observed in Emery-Dreifuss muscular dystrophy patients (EDMD) [1, 2]. In the present study we focus on the evaluation of interactions between MMPs and TIMPs in healthy subjects and in EDMD patients.
Materials and methods: 25 EDMDpatients (connected with lamins AD-EDMD or emerin X-EDMD deficiencies) and 20 aged-matched healthy controls were examined. The serum levels of MMPs and TIMPs were quantified using ELISA sandwich immunoassay procedure. The structural models of MMPs and TIMPs were determined applying molecular modelling methods.
Results: Significant changes of MMPs and TIMPs were noted. The level of MMP-2 was increased in all patients, whereas MMP-9 was increased only in some of these patients. The level of TIMP-1 was either normal or increased, and TIMP-2 was reduced in some EDMD patients. Regarding TIMP-3, it was decreased in all the examined EDMD cases. The structural alignment of TIMP-1, TIMP-2 and the N-terminal domain of TIMP-3 reveals strong structure conservation. Major differences were observed in the length of the sA-sB loop, which interacts with the catalytic domain of MMPs, and in the fragment located before the sD beta strand. The structural model of MMP-9–TIMP-1 complexes, obtained with the homology-based modelling, is presented and discussed. Both complexes were stable during the molecular dynamics simulation (6 ns, with the CHARMM22 force field, without any restraints). TIMP-1 interacts with the activated MMP-9 catalytic domain through its N-terminal and with the hemopexin domain by its C-terminal part. None of the residues conserved in MMPs family occurs in the binding interface.
Conclusions: The presented molecular models explain interactions in the selected complexes of TIMPs and MMPs. The data suggests that the altered balance between some of MMPs and TIMPs in EDMD patients can be considered as risk and prognostic markers.

References:

1. Niebroj-Dobosz I *et al* (2009) *Acta Biochim Pol* **56**: 717-722.
2. Niebroj-Dobosz I *et al* (2014) *Int J Cardiol* **173**: 324-325.

Acknowledgments

These studies were partially supported by the Biocentrum-Ochota project [POIG.02.03.00-00-003/09].

P4.38

Analysis of equilibria in Fisher's geometric model with environmental stress

Michał Startek[1], Arnaud Le Rouzic[2], Anna Gambin[1]

[1]Institute of Informatics, University of Warsaw, Warsaw, Poland; [2]Laboratoire Évolution, Génomes et Spéciation, Centre National de la Recherche Scientifique, France
e-mail: Michał Startek <michal.startek@mimuw.edu.pl>

Fisher's geometric model is a well-established (class of) models in population genetics, wherein the organism's phenotype is represented in a geometric setting, as a vector of real-valued parameters. Selection is based on a fitness function (dependant on the phenotype), and random mutations are modelled by random changes to the organism's phenotype. The population is represented as a probability measure over the space of possible phenotypes, representing the proportion of organisms with given phenotype. The applications of these models range from studying adaptation to an environment, to simulating the evolutionary effects of pleiotropy.
In a mutator model with environmental stress, the organisms are differentiated also by their mutation rate, which in some scenarios is allowed to vary as well. Environmental stress is reflected by moving the optimum of the selection function in each generation.
In our study we assume a Gaussian selection function, as well as Gaussian mutation operator. Exploiting the various properties of the class of Gaussian functions, such as its stability and closure under various operations, allows us to analytically derive the equilibrium probability measure (with respect to the moving optimum) describing the stable state of a population chasing a shifting (with constant speed) phenotypic optimum. Furthermore, convergence of a population described by any probability measure (be it continuous or singular) to the aforementioned equilibrium may be studied and analytically proven. We have performed studies of interdependence of the model's variables and the ability of the modelled population to survive (as even if there is a theoretical equilibrium state it is not realistic for a population which loses a too big fraction of its headcount in each generation to sustain itself).
The selection of Gaussian mutation and selection has allowed to explicitly derive the closed-form formulas for equilibrium state, as well as has enabled the derivation of closed-form exact formulas for various traits of the population, such as genetic variance or average fitness, as a function of the parameters of the model. A sketch of proofs for the aforementioned results shall be presented.

P4.39

DMG-α: computational geometry package for analysis of molecular dynamic simulations

Robert Szczelina[1,2], Krzysztof Murzyn[3]

[1]Jagiellonian University, Faculty of Mathematics and Computer Science, Kraków, Poland; [2]Małopolska Centre of Biotechnology, Department of Bioinformatics, Kraków, Poland; [3]Jagiellonian University, Faculty of Biochemistry, Biophysics and Biotechnology, Department of Computational Biophysics and Bioinformatics, Kraków, Poland
e-mail: Robert Szczelina <robert.szczelina@uj.edu.pl>

The *Power Diagram* (PD) of a set of balls in three dimensions assigns to each of those balls a convex polyhedron, called a *power cell*, that contains all points in the space that are closer to this ball than to any other (with respect to some pseudo-distance function). If two cells share a common face then we call them adjacent and this relation allows to rigorously investigate inner topological structure of the (union of) set of balls. PD can be used to construct data structures and quantities such as Alpha Shapes (to identify pockets and tunnels), Solvent Accessible Surface Area (SASA) and the volume of the union of balls.

Although very useful, PD may be hard to compute in higher dimensions - optimal algorithms for a set of n balls in d-dimensional space needs time proportional to $n^{d/2}$ in the worst case. However, in the context of large, uniformly distributed sets inside a compact subset of R^3 (euclidean three-dimensional space), it was shown that expected running time is proportional to n. Therefore it is feasible to compute PD for sets originated from Molecular Dynamic (MD) simulations, especially for solvated models of biomolecules.

We have recently developed DMG-α library for performing computations of PD and related data structures with stress on algorithms for sets constrained to a simulation box with periodic boundary conditions (PBC) to guarantee that resulting PD and derived data structures are also periodic. The expected computational time is linear and computations may be simply and effectively parallelized. The library is written in C++ but rich Python interface is integral part of the library for easy scripting and extending of existing routines.

To illustrate possible applications of the DMG-α library, we present results of sample analyses which allowed to determine non-trivial geometric properties of two *Escherichia coli*-specific lipids as emerging from molecular dynamics simulations of relevant model bilayers [1].

DMG-α along with all scripts are available for download under terms of BSD license to anyone interested from the project webpage: http://dmgalpha.scirsc.org.

Reference

1. Szczelina R, Murzyn K (2014) DMG-α – a computational geometry library for multi-molecular systems. *J Chem Inf Mod* submitted.

Acknowledgements

Robert Szczelina and Krzysztof Murzyn acknowledge financial support from Polish National Science Centre under grants no. 2011/01/N/ST6/07173 and 2011/01/B/NZ1/00081, respectively.

P4.40

Identification of STAT1 and STAT3 specific inhibitors using comparative virtual screening and docking validation

Małgorzata Szeląg[1], Anna Czerwoniec[3], Joanna Wesoly[2], Hans A. R. Bluyssen[1]

[1]Department of Human Molecular Genetics, [2]Laboratory of High Throughput Technologies, [3]Bioinformatics Laboratory, Institute of Molecular Biology and Biotechnology, Adam Mickiewicz University in Poznan, Poznań, Poland
e-mail: Małgorzata Szeląg <grete@amu.edu.pl>

Signal transducers and activators of transcription (STATs) facilitate action of cytokines, growth factors and pathogens. STAT activation is mediated by a highly conserved SH2 domain, which interacts with phosphotyrosine (pTyr) motifs for specific STAT-receptor contacts and STAT dimerization. The active dimers induce gene transcription in the nucleus by binding to specific DNA-response elements of target genes. Abnormal activation of STAT signaling pathways is implicated in many human diseases, like cancer, inflammation and auto-immunity. Searches for STAT-targeting compounds, exploring the pTyr-SH2 interaction area, yielded many small molecules for STAT3 but sparsely for other STATs. However, many of these inhibitors seem not STAT3-specific, thereby questioning the present modeling and selection strategies of SH2 domain-based STAT inhibitors.

We generated new 3D structure models for all human (h) STATs (1, 2, 3, 4, 5A, 5B and 6) and applied a comparative *in silico* docking strategy to obtain further insight into STAT-SH2 cross-binding specificity of previously identified STAT3 inhibitors. Indeed, by primarily targeting the highly conserved pTyr-SH2 binding pocket the majority of these compounds exhibited similar binding affinity and tendency scores for all STATs. By comparative screening of a natural compound library and using the "comparative STAT3 and STAT1 affinity and binding tendency scores" as a selection criterion, we provided initial proof for the possible existence of STAT3 as well as STAT1-specific inhibitors. *In silico* screening of a multi-million "clean lead" compound library for binding of all STATs, likewise identified potential specific inhibitors for STAT1 and STAT3 after docking validation.

Based on comparative virtual screening and docking validation, we developed a novel STAT inhibitor screening tool that allows identification of specific STAT1 and STAT3 inhibitory compounds. This could increase our understanding of the functional role of these STATs in different diseases and benefit the clinical need for more drugable STAT inhibitors with high specificity, potency and excellent bioavailability.

Acknowledgements

This research was supported by grants: UMO-2012/07/B/NZ1/02710 (to HB), UMO-2012/07/N/NZ2/01359 (to MS) from National Science Center Poland and No 128 from the Poznan Supercomputer Center (PCSS) (to MS). This research was supported in part by PL-Grid Infrastructure.

P4.41

Does the sickle-cell mutation affect the DNA conductivity? Redox non-innocent role of riboflavin: experimental and computational studies

Diana Toczydłowska[1], Łukasz Górski[2], Piotr Zarzycki[1]

[1]Institute of Physical Chemistry of the Polish Academy of Sciences, Warsaw, Poland; [2]Warsaw University of Technology, Department of Microbioanalytics, Warsaw, Poland
e-mail: Diana Toczydłowska <dtoczydlowska@ichf.edu.pl>

The change of one base pair in a DNA sequence (point mutation) is a common genesis of many genetic diseases (e.g. mucoviscidosis, sickle-cell anemia, color blindness). One of the most widespread PM is associated with the sickle-cell anemia (SCA). Surprisingly, the single A-T to T-A replacement in the 6 codon of β-hemoglobin encoding sequence results in a dramatic hemoglobin misfolding. This results in the overall red-blood cell shape-deformation (sickle shape), which increases hydrophobicity of hemoglobin that leads to aggregation of proteins [1] SCA is responsible for the death of over 2.5 hundred thousand of African American population [2].
On the other hand, since DNA is capable of conducting current via the hopping or tunneling mechanisms [3], it is of great interest to examine the effect of SCA mutation on its conductivity. In particular, SCA detection on an early stage of child development in a noninvasive way (e.g., biosensors with nontoxic hybridization indicator, for instance riboflavin) is highly demanded.
In our work, we selected 15 nucleotides ss-DNA chain from hemoglobin sequence (HBB) (GenBank accession number Gu324922.1) to prepare the SCA biosensor. We observed that conductivity of DNA is sensitive to the SCA mutation. Our experimental studies are accompanied by the molecular dynamics simulations and ab-initio calculations, which shed the light on nature of the conductivity signature of the SCA mutated DNA.

References:

1. De Llano JJ *et al* (1994) *Protein Science* **3**: 1206–1212.
2. Lanzkron S *et al* (2013) *Public Health Rep* **128**: 110–116.
3. Porath D *et al* (2000) *Nature* **403**: 635–638.

P4.42

RNPDock: a server for protein-RNA complexes structure refinement

Irina Tuszynska[1], Marcin Magnus[1], Janusz M Bujnicki[1,2]

[1]International Institute of Molecular and Cell Biology, Laboratory of Bioinformatics and Protein Engineering, Warsaw, Poland; [2]Institute of Molecular Biology and Biotechnology, Bioinformatics Laboratory, Adam Mickiewicz University, Poznań, Poland
e-mail: Irina Tuszyńska <irena@genesilico.pl>

Protein-nucleic acid interactions play an important role in the life of every cell. Methods that are able to predict protein-RNA complexes are needed to understand the principles of protein-RNA recognition and next to design new RNA-binding proteins. Most protein-RNA docking methods carry out docking of rigid structures of macromolecules. These limitations often lead to sub-optimal sampling of the complex conformational space, resulting in protein-RNA model structures with native interaction sites but not optimal relative position of the components.
We provide an effective method of the conformational space sampling within an interaction area of complex components – RNPDock (http://iimcb.genesilico.pl/RNPdock/).Our method uses Monte Carlo algorithm combined with DARS-RNP statistical potential (Tuszynska *et al*., 2011, *BMC Bioinformatics* **12**: 348). For more effective conformational space sampling, and quickly bringing the system to the global energy minimum, the simulated annealing algorithm, which gradually reduces the temperature of the system, was implemented. As a result, RNPDock finds the structure of protein-RNA complex with the best score.
The method was tested on the unbound docking set from Tuszynska *et al.* (2011, *BMC Bioinformatics* **12**: 348). Started structures were taken from the first biggest cluster of the best scored structures. Their average RMSD value was around 10 Å. The average RMSD improvement, after RNPDock running, was around 3 Å.

P4.43

Structural rearrangement in coiled-coil region during fibrinogen to fibrin transformation

Lesia Urvant, Nikolai Pydiura

Palladin Institute of Biochemistry, National Academy of Sciences of Ukraine, Kyiv, Ukraine
e-mail: Lesia Urvant <lpurvant@gmail.com>

Fibrinogen (Fg) is a polyfunctional protein of blood coagulation system. Fibrin, formed by cleavage of the pairs of fibrinopeptides A and B from Fg by the thrombin, polymerizes in a two steps – protofibrils formation and their lateral association. Previously by using fibrin-specific monoclonal antibody I-3c and synthetic peptides Bβ121-138 and Bβ109-126 we found a site Bβ126-135 in coiled-coil connector which provides lateral association of protofibrils. Computer modeling in Modeller 9v10 software showed that site Bβ126-135 is exposed in fibrin molecule following changing of accessibility of these amino acid residues to solvent and partial uncoiling of a-coil. Amino acid sequence Bβ126-135 is the structural component of the part of the coiled-coil connector. In this part of the coiled-coil Bβ126-135 connects with g69-77 and Aa91-103.
In this work we testify the influence of the synthetic peptide which correspond to the amino acid residues of the fibrin(ogen) molecule g69NPDESSKPN78 on the fibrin polymerization process. Turbidity analysis and transmission electron microscopy showed that synthetic peptide g69-77 inhibits the second step of fibrin polymerization – the protofibrils lateral association. Structural dynamics of peptide g69-77 indicate that in an aqueous solution it has a form of a loop, stabilized by hydrogen bonds. Such conformation of this peptide is necessary for providing its inhibitory activity.
We suggest that a possible mechanism of inhibition effect of peptide γ69-77 is a competitive interaction with Aa91-103 and Bβ126-135 a-coils, and, as a result – disturbance of the coiled-coil structure in this part of the molecule. According to this suggestion we assume that site Bβ126-135 and g69-77 are required for increasing flexibility of this part of coiled-coil but not for the direct interaction between protofibrils.

P4.44

Dependency between correlated mutations and occurrence of single-nucleotide polymorphism

Pawel P. Wozniak[1], Gert Vriend[2], Malgorzata Kotulska[1]

[1]Institute of Biomedical Engineering and Instrumentation, Wrocław University of Technology, Wrocław Poland; [2]Centre for Molecular and Biomolecular Informatics, Radboud University Nijmegen Medical Centre, The Netherlands
e-mail: Paweł Woźniak <pawel.p.wozniak@pwr.edu.pl>

One of the most significant types of genetic variations is single-nucleotide polymorphism (SNP). SNPs are the most useful, and frequently applied markers in genetic science. However, the characteristic of SNPs occurrence has not been fully examined yet. A proper understanding of the relation between SNPs and protein sequence features can help in automation of sequence analysis and provide more information about the characteristic of SNPs. One of these features might be the importance of a single amino acid for the stability of the whole protein structure. This information can be obtained e.g. from the analysis of protein contact sites. Algorithms based on the correlated mutations are one of the most popular methods used in the protein contact sites prediction. The Direct-Coupling Analysis (DCA) algorithm has been reported as particularly effective, recently. It is an enhancement of the older Correlated Mutation Analysis (CMA) algorithm, which predicts protein contacts based on the information about correlated mutations occurring in the multiple sequence alignment of the query protein [1]. The aim of this study is to analyze whether there is any dependency between the results of DCA and CMA algorithms and the SNPs occurrence.
The DCA and CMA algorithms were implemented in Java 1.7.0. A dependency between the results of these algorithms and SNPs occurrence was analyzed for the group of 25 human proteins possessing the highest number of missense SNPs. This dataset was obtained from the Exome Sequencing Project [2]. Furthermore, the frequency of the SNPs occurrence was investigated for different amino acid types.
The study shows that it is possible to estimate the ranges of the values obtained from the DCA and CMA algorithms which are usually taken by the amino acids at SNP positions. However, this is not enough to unambiguously point SNP amino acids and additional sources of information are needed. On the other hand, the results prove that the frequency of the SNPs occurrence within different amino acid types is strongly related to the occurrence of so called CpG islands.

References:

1. Morcos F *et al* (2011) *Proc Natl Acad Sci USA* **108**: E1293-301.
2. Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (URL: http://evs.gs.washington.edu/EVS/ retrieved 12.09.2012).

P4.45

DNA metabarcoding of the kombucha multi-microbial community

Iryna Zaets[1], Olga Kukharenko[1], Olga Podolich[1], Oleg Reva[2], Natalia Kozyrovska[1]

[1]Institute of Molecular Biology & Genetics of National Academy of Sciences of Ukraine, Laboratory of Microbial Ecology, Ukraine; [2]University of Pretoria, Department of Biochemistry, Bioinformatics and Computational Biology Unit, South Africa
e-mail: Iryna Zaiets <zkora@ukr.net>

Objective: DNA metabarcoding is a means of the identification of living organisms and discrimination between them through the comparison of short standardized DNA sequences - barcodes. Taking into account that the majority of microbial species cannot be cultured and hard to be distinguished by any microbiological methods, the species profiling by high-throughput DNA pyrosequencing is the best of known tools to control microbial communities. Pyrosequencing-based barcoding of microbial communities originated from fermented foods is essential for the deciphering a successful design of probiotic and synbiotic products. In this work, we applied pyrosequencing of 16S rDNA and ITS amplicons obtained from bacterial and yeast species, composing the Ukrainian ecotype of kombucha culture (UEKC), also known as the fermented tea drink.
Methods: 454 pyrosequencing of amplicons used as barcodes for species profiling of UEKC was used. DNA samples were isolated from the liquid phase and biofilm fraction of 14-day kombucha culture. The obtained DNA reads were aligned locally by BLASTN against combined NCBI 16S Microbial and Greengenes16S databases for identification of 16S rDNA reads and against the NCBI nt-database for identification of ITS reads. The BLASTN results were merged and visualized by MEGAN 4.67.4. Also the BLASTN output files were searched by an in-house BioPython-based script to retrieve the statistics of the top scored hits for all reads.
Results: DNA metabarcoding exhibited a more complex kombucha culture structure than it was known before: it was comprised by two bacterial (*Proteobacteria, Firmicutes*) and yeast (*Ascomycota*) phyla, as well as several unknown pro- and eukaryotic microorganisms. Kombucha is quite flexible and variable community, but it produces a stable core microbiome in different conditions of cultivation. The core community includes acetobacteria of two genera (*Komagateibacter, Gluconobacter*) and yeast, which mainly belong to *Brettanomyces/Dekkera, Pichia*, and *Candida* genera.
Conclusions: Based on these data, the precise species inventory of microbial association will get us insight on the functionality of this community as whole entity. DNA profiling by pyrosequencing is a new powerful tool for characterizing dynamic changes in probiotic communities and in the gut microbiota treated with probiotics. This knowledge may aid in improving of probiotic development and administration for prophylaxis of human diseases.

P4.46

Integrative data analysis for DNA damage repair genes related to p53

Joanna Zyla[1], Christophe Badie[2], Ghazi Alsbeih[3], Joanna Polanska[1]

[1]Silesian University of Technology, Institute of Automatic Control, Gliwice, Poland; [2]Public Health England, Chilton, Didcot, United Kingdom; [3]King Faisal Specialist Hospital & Research Centre, Riyadh, Kingdom of Saudi Arabia
e-mail: Joanna Zyla <joanna.zyla@polsl.pl>

Aim: Aim of this study is to propose data analysis method based on integrative analysis which could be used for analysis many genes in one time. In this research proposed methodology is used for analysis genes which appear in DNA damage repair pathway.
Materials and methods: The population under investigation is composed of 44 unrelated, healthy individuals, where for each, two type of data were collected. First was the result of genotyping of 567,095 polymorphisms (SNP) by Axiom platform, the second one includes GADD45 and SESN1 qPCR expressions measured after the irradiation of 2Gy and in normal conditions (0Gy). Additionally the γH2AX test on radiosensitivity was performed. All of three measured genes are related with p53 branch of DNA damage repair. For every gene, statistical significance were calculated by the most powerful test to each SNP (ANOVA, t-test or *Mann-Whitney*). After, the 3 groups of p-value were integrated by Z transformation method of p-value combining. Finally, obtained SNPs were investigated by their transcriptomic locations, KEGG and Gene Ontology (GO) to show level of biological significance.
Results: During the first step, the analysis of genotype-phenotype interaction between every SNP gene expression was investigated. Having a set of three of p-values obtained for each gene, allow to performed p-value combing by Z transformation method. Using the significant level α=0.05 number of candidate SNPs for integrative method are equal 43 188 and 46 374 (for 0Gy and FCH respectively). While the number of intersect SNPs through all three genes are equal 1920 and 1317 (for 0Gy and FCH, respectively) — restricted method. Additionally the level of transcriptomic SNPs stay on the same level 38% to 42% (FCH — restricted and integrative method), which gives the same level of biological significance. Additionally, by testing the overrepresentation of signaling pathways from KEGG, there was no significant pathway for restricted method. While for integration e.g. for FCH the analysis show overrepresentation for such pathways like: "pathway in cancer" (p-value-0.0001) and "MAPK signaling pathway" (p-value-0.004). Both of them are highly relevant to investigated problem. What is more for integrative method number of overrepresented GO increase twice time to each class.
Conclusions: The data integration technique couldsignificantly increase the number of candidate polymorphisms for investigated problem. In parallel, integrative method increase the level of significance for biological information. Group of candidate SNPs for DNA damage repair genes related to p53, was obtained.

Acknowledgement

The work was financially supported by NCN grant HARMONIA 4 register number 2013/08/M/ST6/00924 (JP, CB), SUT- BKM/524/Rau1/2014/t.16 (JZ). Additionally, JZ is holder of scholarship DoktoRis – Scholarship program for Innovative Silesia. Calculations were carried out using infrastructure of GeCONiI (POIG.02.03.01-24-099/13).

P4.47

Reduction of gene signatures based on fusion of expression and ontology information

Wojciech Labaj[1], Andrzej Polanski[1]

[1]Institute of Informatics, Silesian University of Technology, Gliwice, Poland
e-mail: Wojciech Łabaj <wojciech.labaj@polsl.pl>

Gene signatures are lists of genes used for summarizing high-throughput gene expression experiments. There are routines for obtaining and analyzing gene signatures in molecular biology researches, including statistical testing with false discovery corrections and annotations by gene ontology (GO) keywords. Despite established routines there are still challenges in efficient applications of gene signatures, which include instability of composition of gene signatures, problems in defining sizes (numbers of genes) of gene signatures and possible unreliability of results of inference based on gene signatures. Therefore there are many efforts towards improving algorithms for constructions of gene signatures. In this paper we are introducing a methodology of reduction of a gene signature corresponding to a gene expression measurement experiment, based on the fusion of information coming from statistical testing of differential expression genes and information resulting from statistical testing of enrichment for GO terms.

On the basis of the DNA microarray datasets we are demonstrating that the proposed algorithm for fusion of expression and ontology information leads to improvement of the composition of gene signatures. We are illustrating our algorithm for integrative construction of gene signatures and we are comparing it with other gene annotation algorithms on the basis of the results of DNA microarray experiments on astrocytic brain tumors. Astrocytic brain tumors are cancers of primary central nervous system (CNS), which develop from astrocytes and are most common glial tumors. They can be divided into two groups according to the way of growth, diffuse or localized. In our study we are focusing on the astrocytic brain tumors with the diffuse growth, which give poorer prognosis and are assigned to a higher grade according to the World Health Organization (WHO). We are further confining the research to the two most common tumors from this group, namely anaplastic astrocytoma (AA) and glioblastoma multiforme (GBM). Additionally, there are two different forms of the GBM, primary GBM arising de novo and secondary GBM arising from lower grade diffuse astrocytoma, anaplastic astrocytoma. Primary and secondary glioblastomas are histologically indistinguishable, except the facts that the frequency of extensive necrosis is higher for the primary GBM and the frequency of oligodendroglioma components is higher for the secondary GBM. Similar histopathology of glioblastomas may be due to similarity of genetic alterations behind their growths.

On the basis of the biological and clinical characteristics of astrocytoma and primary and secondary glioblastomas, it seems an interesting issue to design an experiment based on the high throughput techniques of molecular biology, aimed at visualization of differences/similarities between these cancers. We have searched the Gene Expression Omnibus (GEO) database for gene expression profiles corresponding to the above-mentioned tumors and their comparisons to the normal tissues. The gene expression datasets found in GEO were obtained in studies, which relied on comparisons of the three types of brain tumors, AA, primary GBM (GBM.P) and secondary GBM (GBM.S) with the normal brain tissues (NBT).

We are comparing different methodologies of reduction of gene signatures for AA, GBM.P and GBM.S experimental data. The classical annotation routine gives results consistent to biological backgrounds of diseases, with numerous GO terms shared by GBM.P and GBM.S and rather few GO terms shared between AA and GBM. Interestingly this relation between numbers of GO terms is not preserved by majority of gene signature reduction algorithms from the literature. We show that the proposed method of gene signatures reduction based on fusion of two p-values both limits the number of genes and preserves the structure of GO terms shared between AA, GBM.P and GBM.S.