Session 20: Bioinformatics

Lectures

L20.1

Combining wet-lab approaches and bio-informatics for the understanding of fungal non-self recognition and immunity

Sven J. Saupe[1], Witold Dyrka[2], Matthieu Paoletti[1], Asen Daskalov[3], Corinne Clavé[1], Fred Ness[1]

[1]Institut de Biochimie et de Génétique Cellulaire, CNRS, Université de Bordeaux, Bordeaux, France; [2]Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wroclaw University of Technology, 50-370 Wroclaw, Poland; [3]Department of Plant and Microbial Biology, University of California, Berkeley 341A Koshland Hall, Berkeley, CA 94720, USA

Sven J. Saupe <sven.saupe@ibgc.cnrs.fr>

The recognition of non-self is essential for all cellular life forms. While these processes have been essentially studied in metazoans and higher plants, they operate also in other branches of the eukaryotic tree. We aim at a better understanding of the immune and non-self recognition processes in the fungal branch. We found in particular that like plants and animals, fungi display large repertoires of Nod-like receptors (NLRs) some of which at least function in the control of programmed cell death in response to the detection of non-self. We will attempt to describe how the question of the immune response in fungi is approached by a combination of wet-lab and bio-informatic approaches. In particular, we will describe the mechanism of diversification of NLRs in fungi and the role of prion amyloid signalling motifs in their activity.

L20.2

HOPE for the future? Bioinformatics to the rescue!

Gert Vriend

Radboud University Medical Centre, Centre for Molecular and Biomolecular Informatics (CMBI), Nijmegen, Netherlands

Gert Vriend <Gerrit.Vriend@radboudumc.nl>

Solving a problem often starts with having a problem, and describing it. There probably doesn't exist a bigger problem than having a child with a birth-defect. In our hospital medics deal with these problems. Sometimes it is simple; if a child has Down syndrome, or some other well-characterized birth-defect, then thee doctor can tell the unfortunate parents what to do, what to expect, etc. However, there are 18 000 000 000 possible ways to alter a genome once, and there are about 10 thousand described genetic disorders. Needless to say that our genetic doctors very often have to tell parents "We have no idea what is wrong with your child, but we will work on it". They then try to find the genomic cause for the phenomic problem. Most often that is one of those 18 billion possible variants, and once they found that, what then? That is where bioinformaticians come to the rescue. Once we have homology modelled the structure in which the mutation was observed, we can start thinking about the molecular phenotype underlying the disease phenotype. Unfortunately, the number of bioinformaticians who can look at a protein structure and then actually see something is way, way too small. And that is why we made HOPE, the Siri of human genetics. That required a lot of AI (or whatever it is you call it when you have to cope with text written by a medic), and some hard-core informatics. But at the end there was HOPE and hope.

L20.3

Identification of informative variables: a tool for knowledge discovery in life sciences

Witold R. Rudnicki[1,2], Krzysztof Mnich[3], Szymon Migacz[2], Paweł Tabaszewski[2], Andrzej Sułecki[2], Aneta Polewko-Klim[1], Wojciech Lesiński[1], Agnieszka Golińska[1]

[1]Institute of Informatics, University of Białystok, Białystok, Poland; [2]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Poland; [3]Computational Centre, University of Białystok, Białystok, Poland
Witold R. Rudnicki <W.Rudnicki@uwb.edu.pl>

The goal of scientific research is to build a possibly accurate model of reality that allows for understanding and possibly also predicting studied phenomena. To this end molecular biologists for many years have applied the very successful strategy: first state your research hypothesis, then design and perform the experiment that would either confirm or refute it. Modern experimental technologies, such as analysis of gene expression, next generation sequencing, proteomics, allow for another approach. One collects immense amounts of data and then explores the data for search of interesting phenomena using statistical and machine learning methods.

In many cases the main difficulty is shifted from experiment design and data generation to the analysis of the data. There are multiple problems with this approach, however. In particular one has to take into account that multiple (practically infinite number) hypotheses are tested in parallel and the number of variables in the models are usually much higher than number of objects under scrutiny.

Identification of variables that carry information about studied phenomenon is important step in the analysis. In many cases it is sufficient to find a small subset of variables with strongest association with the observed effect that are sufficient for building a predictive model. Identification of a set of marker variables, which can be used to discriminate between two types of cells, is an example of such application. However, this approach is not always sufficient. In many cases only a small part of variables is truly responsible for the observed phenomena, the other are part of the response to the original causal variables. Moreover, the strength of association may be much higher in these response variables, than the causal variable – for example the relatively small change of expression level of some regulatory protein may result in turning on or off expression of hundreds of proteins in response. Only identification of the causal genes can lead to understanding of the phenomena under scrutiny.

We present several approaches for identification of all variables that carry information in the information system. They are based on the notions of weak and strong relevance of variables and use mix of machine learning, statistics and information theory. The approach is general and can be applied in various fields of modern molecular biology.

Posters

P20.1

Hypereosinophilic syndrome – modeling and analysis using the Petri Net theory

Alexander Antkowiak[1], Dorota Formanowicz[2], Piotr Formanowicz[1,3]

[1]Institute of Computing Science, Poznan University of Technology, Poznań, Poland; [2]Department of Clinical Biochemistry and Laboratory Medicine, Poznan University of Medical Sciences, Poznań, Poland; [3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland
Alexander Antkowiak <alexander.antkowiak@cs.put.poznan.pl>

Many diseases are known for their increased mortality; however some of these diseases are treatable if they are early detected. One of such diseases is hypereosinophilic syndrome (HES). HES is characterized by persistent eosinophilia that is associated with damage to multiple organs over time. The extent of organ damage depends mostly on how early the disease is detected. An early diagnosis gives a chance to get this disease under control and to longer lifespan of the patients. Without treatment this disease is fatal. The question we can ask is if there is anything we can do to improve the treatment? Certainly, we need a better understanding of the processes that underline this disease. For this purpose, we have built their model expressed in the language of Petri Net theory. It could give a more comprehensive understanding of this disorder and in the result it could give more answers in dealing with this disease.

Acknowledgements:

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

P20.2

Automated brain tumour segmentation with the use of MiMSeg algorithm. Evaluation on BRATS 2013 challenge dataset

Franciszek Binczyk[1], Michael Goetz[3], Christian Weber[3], Bram Stjelties[4], Klaus Meier-Hein[3], Hans-Peter Meinzer[5], Barbara Bobek-Billewicz[2], Rafal Tarnawski[2], Joanna Polanska[1]

[1]Data Mining Group, The Silesian University of Technology, Gliwice, Poland; [2]Center of Oncology - Maria Sklodowska-Curie Memorial Institute, Branch in Gliwice, Poland; [3]Junior Group Medical Image Computing, German Cancer Research Center, Heidelberg, Germany; [4]Department of Radiology, University Hospital Basel, Switzerland; [5]Division Medical and Biological Informatics, German Cancer Research Center, Heidelberg, Germany

Franciszek Binczyk <franciszek.e.binczyk@polsl.pl>

Aim: The automated detection of tumour on medical images is still not solved. The accuracy of automated detection in comparison with manual experts varies from ~70% to ~87%. In this work a MiMSeg algorithm, originally developed by authors for the diffusion weighted imaging data is adjusted to operate on T1 and T2 FLAIR images.
Material and methods: The MiMSeg first step is decomposition of training images into the Gaussian mixture model. Next, all obtained components are clustered with the use of Dunn's index driven k – means algorithm in the space of mixture model component parameters. As a last step the conditional probability that certain value (T1 of FLAIR) belongs to the certain cluster and a cut – off value is estimated.
Data: A BRATS 2013 set consisting of T1 and T2 FLAIR images (separate: training and validation) has been used. Each set contains images for two types of tumour: low and high grade what allows for tumour type dependant analysis. The high grade set consist of 2196 (training) and 272 (validation) for both T1 and FLAIR. Low grade tumour, set contains of 1032 (training) and 284 (validation) images. All images are normalized and co-registered. The only processing performed was cerebrospinal fluid filtration.
Results: The MiMSeg was trained for low and high grade tumours, both for T1 and FLAIR sequences separately. The mean DSC values obtained on validation set of high grade tumour were equal to: 81.15% for FLAIR and 60.88% for T1. As for low grade tumour the result were 83.23% mean DSC for FLAIR automated detection and 65.36% for T1.
Discussion and conclusion: As suspected the use of FLAIR images resulted in better accuracy of automated grand tumour volume detection than T1 images. It is caused by the fact that T1 show mostly the destruction of blood-brain barrier (including but not only, necrosis) and the other effects of tumour presence are non detectable. When considering only necrosis T1 driven detection (grand tumour volume subregions are available for BRATS data), the results rises to the 86.54% for high and 88.06% for low grade tumour. The obtained results prove MiMSeg to be the universal algorithm that can operate on variety of medical images. In the future authors will combine automated annotation on different medical images into a multimodal technique.

Acknowledgments:

The study was financed by SUT grant: 02/010/BKM16/0047/t. 25. Calculations were carried out on IT infrastructure of GeCONiI Upper Silesian Centre for Computational Science and Engineering (POIG.02.02.01-24- 099/13).

P20.3

AmyloGram: analysis and prediction of amyloids using n-grams

Michał Burdukiewicz[1], Piotr Sobczyk[2], Stefan Rödiger[3], Anna Duda-Madej[4], Paweł Mackiewicz[1], Małgorzata Kotulska[5]

[1]University of Wrocław, Wrocław, Department of Genomics, Poland, [2]Wrocław University of Science and Technology, Faculty of Pure and AppliedMathematics, Wrocław, Poland, [3]Brandenburg University of Technology Cottbus-Senftenberg, Institute of Biotechnology, [4]Wroclaw Medical University, Department of Microbiology, Wrocław, Poland; [5]Wrocław University of Science and Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław, Poland

Michał Burdukiewicz <michalburdukiewicz@gmail.com>

Amyloids are proteins associated with the number of clinical disorders (e.g., Alzheimer's, Creutzfeldt-Jakob's and Huntington's diseases). Despite their diversity, all amyloid proteins can undergo aggregation initiated by 6- to 15-residue segments called hot spots. Henceforth, amyloids form unique, zipper-like β-structures, which are often harmful. To find the patterns defining the hot spots, we developed our novel predictor of amyloidogenicity AmyloGram, based on random forests. We trained it using short motifs (n-grams) extracted from amyloid and non-amyloid peptides collected in the AmyLoad database.
Peptide data were represented by various amino acid physicochemical properties. We tested 524 284 random forest predictors, each employing reduced amino acid alphabet based on a different combination of the physicochemical properties of residues. As a result, we identified the reduced alphabet providing the best discrimination between amyloids and non-amyloids, which was based on the hydrophobicity index, polarizability parameter, β-sheet propensity and average flexibility. Three first features are well-known factors in amyloidogenicity, but the role of the last one in this process was previously unknown. Most of the predictors based on reduced amino acid alphabet outperformed a random forest trained on the full amino acid alphabet confirming our assumption on the role of more general amino acid properties. During analysis we also found 65 n-grams that are most relevant to the discrimination of amyloid and non-amyloid sequences, 15 motifs were independently confirmed experimentally elsewhere.
The best-performing predictor, AmyloGram, was benchmarked against the most popular tools for amyloid peptides detection using an external dataset. Our software obtained the highest values of performance measures (Area Under the Curve: 0.90, Matthews correlation coefficient: 0.63).
The n-gram analysis not only confirmed that amyloidogenicity depends on the general physicochemical properties of proteins, but also revealed which features are the most relevant to the initiation of amyloid aggregation. In addition, our framework identified amyloidogenicity-related amino acid motifs, which were partially confirmed experimentally. Aside from creation of the interpretative model of amyloidogenicity, we also established the accurate predictor of amyloids, AmyloGram, which is available as a web-server: www.smorfland.uni.wroc.pl/amylogram/.

P20.4

Should we reconsider the chromosomal gene movement *via* retroposition?

Joanna Ciomborowska, Michal Kabza, Izabela Makalowska

Faculty of Biology, Department of Bioinformatics, Adam Mickiewicz University in Poznan, Poznań, Poland
Joanna Ciomborowska <joannac@amu.edu.pl>

Retrocopies are generated through reverse transcription of multi-exonic parental gene transcripts and insertion of originated cDNA into the genome. They usually lack introns and regulatory elements and most of them become pseudogenes. However, it is also known that some fraction of them may gain functionality and play important roles in genomes. One of the hallmarks of retroposition is duplication of the gene in a new genomic location. Most of previous studies showed a strong tendency for movement from chromosome X to autosomes.

Here we examined, using bioinformatics and statistics methods, chromosomal gene movement via retroposition for 15 mammalian species collected in RetrogeneDB. We analysed the movement for chromosome X and autosomes to test hypothesis of the excess of retroposition from X to autosomes on bigger scale and also for each chromosome to identify the best „donors" and „acceptors" of retrocopies.

Our results revealed that chromosomal gene movement via retroposition seems to be more random than it was previously thought. Moreover there is a significant difference in chromosomal gene movement pattern between all and expressed retrocopies. Considering expression patterns we found out also that retrocopies which moved out of X are more often functional so those genes are most probably under positive selection.

Summing up, thanks to our approach, we can follow movement of retrocopies in details and observe differences related to expression pattern. Our results present an interesting perspective and will help to understand retroposition and it's evolutionary consequences better.

P20.5

Quantiprot – a Python package for quantitative analysis of protein sequences

Bogumił M. Konopka, Witold Dyrka

Wroclaw University of Technology, Faculty of Fundamental Problems of Technology, Department of Biomedical Engineering, Wrocław, Poland
Witold Dyrka <witold.dyrka@pwr.edu.pl>

Protein sequence analysis is the fundamental task of bioinformatics. The field is dominated by tools based on local and global alignments and substitution matrices - for good reasons and with unquestionable success. However, there are significant applications where alternative approaches are necessary. For example, alignment-based methods are not suitable for analysis of very divergent and/or large sets of sequences. A viable alternative can be provided by methods of quantitative characterization, often originated from literary linguistics and dynamical systems modeling, which assign vectors of numerical values to entire sequences or their fragments. A major advantage of the approach is that quantitative properties define a multidimensional feature space, where sequences can be related to each other and differences can be interpreted. Despite promising results, only few methods for quantitative analysis have been made available as bioinformatic packages or web services. Nowadays, with enormous amount of data available in public and restricted databases, the need for such tools is evident. We address it by developing Quantiprot - a software package in Python, which provides a simple and consistent interface to multiple methods for quantitative characterization of protein sequences. The package can be used to calculate dozens of characteristics for entire sequences or within sliding windows. The characteristics can be computed directly from amino acid sequences or using physico-chemical properties of amino acids mapped onto sequences. In addition to basic measures, Quantiprot performs quantitative analysis of recurrence and determinism in a sequence, calculates frequency of n-grams and computes the Zipf's law coefficient. We propose three main applications of the package. First, quantitative characteristics can be used in alignment-free similarity searches, and in clustering of large and/or divergent sequence sets. Second, a feature space defined by quantitative characteristics can be useful in comparative studies of protein families or organisms. Third, the feature space can be also very useful for evaluating generative models, where quantitative characteristics of sequences generated by a model can be compared to characteristics of actually observed sequences.

P20.6

Systematic analysis of somatic driver mutations in glioblastoma multiforme TCGA datasets

Mateusz Garbulowski[1,2], Andrzej Polanski[1]

[1]Institute of Computer Science, Silesian University of Technology, Gliwice, Poland; [2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Mateusz Garbulowski <mateusz.garbulowski@polsl.pl>

Next-generation DNA sequencing methods applied to cancer tissues material provide to discover the processes of tumor evolution. The comparison of tumor and normal samples allows to find somatic mutations that are related to tumor growth. Detection and analysis of the clonal structure of tumor cells plays very important role in prediction the cancer expansion, which is described by the clonal theory. Furthermore, the detection of the driver mutations, that are involved in cancerogenesis is very crucial step in the analysis of tumor clonal expansion.

This study was aimed at developing the techniques oriented towards the detection of driver and passenger mutations extracted from whole exome sequencing (WES) experiments of human glioblastoma multiforme (GBM) acquired from The Cancer Genomic Atlas (TCGA) resources. We have created a pipeline of WES data analysis of GBM cancer cells, leading to discover lists of driver and passenger somatic mutations. We decompose the received histograms of measured variant allele frequency (VAF) into Gaussian mixture of probability distributions and we estimate the VAF ranges between clonal and subclonal population of somatic mutations and then we perform the analysis of age status (early or late) of driver mutations. We have also checked the influence of the driver mutations in the survival of patients. In addition, we calculated the number of driver events across the chromosomes for all collected patients.

The results of this project show that there is a list of driver mutations, that are specific for glioblastoma multiforme and some of them are very early events and may influent into patient survival. Moreover, some of the chromosomes accumulate more mutations than others.

Acknowledgements:

This work has been supported by the Silesian University of Technology with the BKM: 02/020/BKM15/0062.

P20.7

Bioinformatic protocol for epitope prediction and its application in analysis of laboratory data for exemplary bacteriophage proteins

Marek Harhala, Katarzyna Hodyra-Stefaniak, Krystyna Dąbrowska

Institute of Immunology and Experimental Therapy Polish Academy of Sciences, Bacteriophage Laboratory, Weigla 12, 53-114 Wrocław, Poland

Marek Harhala <marek.harhala@iitd.pan.wroc.pl>

Vaccine design, phage therapy, protein medicines are medical approaches that require good understanding of immune response. Epitopes (sites recognized by immune system within antigens) determine the scale and the type of immune response. Without recognizing epitopes in a molecule rational planning of vaccines, therapeutic proteins, etc. cannot be applied. It is not enough to find candidate epitopes, but effective application of immune-related approach requires insight into differences between similar epitopes.

Bioinformatic analysis has emerged as a vital point of immunological research. Incorporating modern knowledge, databases, machine learning and computer simulation allowed for creation of new protocols that yield promising candidates for research and help to understand immune responses. Trials to create software&protocols for such procedure are considered partially successful due to problems like: different databases used for machine learning, incomplete knowledge about spatial conformation of epitopes, various scoring systems, lack of knowledge about an effect of single aminoacid (for protein) mismatch on the entire epitope, difficulties in calculation of strength of epitope-immune system interactions.

We propose new protocols to predict&analyse antigenic epitopes in proteins. We used immunological software and framework like FRED2, EPMeta (that combines EPCES, EPSVR, DiscoTope, BEPro), CBTOPE, Abcpred, Bepipred, Expitope, FDR4, IFNepitope, etc. We applied nine exemplary proteins of three phages (Pseudomonas phages: vB_PaeM_CEB_DP1, LMA2 and F8). Results of bioinformatic analysis were compared to experimental data. As the result, we identified 9 candidate epitopes that correspond to immune response profile observed in laboratory experiments.

Acknowledgements:

This work was supported by National Science Centre grants UMO-2015/18/M/NZ6/00412 and UMO-2013/08/M/NZ6/01022.

P20.8

Loss and expression of retrogenes among human populations

Michał Kabza[1], Magdalena Kubiak[1], Agnieszka Danek[2], Wojciech Rosikiewicz[1], Sebastian Deorowicz[2], Andrzej Polański[2], Izabela Makałowska[1]

[1]Department of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University in Poznan, Poland; [2]Institute of Informatics, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland

Magdalena Kubiak <magdalena.kubiak@amu.edu.pl>

Retrogenes are RNA-based genes' copies resulted from reverse transcription of mature mRNA and reintegration of cDNA into a new genomic location. For a long time, this sequences have been considered as non-functional pseudogenes, but recent studies indicate that many of them are expressed in various types of tissues across animal and plant species. Moreover, retogenes are shown also as a driving force of evolution and represent considerable source of polymorphisms between individuals.

In our research we used data from the 1000 Genomes Project to analyze variation of retrogenes presence or absence in human populations, as well as RNA-Seq data to examine their expression level. Furthermore, we performed comparative analysis across fourteen Eutherian species in order to identify retrocopy orthologs. We were able to identify 193 retroduplication variations, from which the majority has not been reported previously. Most of them resulted from the retrocopy deletion. Additionally, we confirmed expression of 11 retrogenes classified as conserved and deleted in some populations. Their expression, level of conservation and low rate of deletion may suggest some functionality. Moreover, we searched for differences in retrocopy expression levels between populations and as a result we detected 9 retrogenes that undergo statistically significant differential expression.

In addition, we developed a novel bioinformatic approach to detect retrogenes not annotated in the reference human genome. We experimentally confirmed the existence of 3 novel (not present in the reference genome) retrocopies and their deletion in some individuals from panel of 17 human genomes supplied by the 1000 Genome Project.

Our findings suggest that retroduplication variations can provide great insight into ongoing evolutionary processes and shed a new light on the determinants of inter-population variation.

P20.9

PyRosetta energy terms as indicators for protein mirror models

Monika Kurczyńska, Bogumił M. Konopka, Małgorzata Kotulska

Wroclaw University of Science and Technology, Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Poland

Monika Kurczyńska <monika.kurczynska@pwr.edu.pl>

In May 2016 the number of protein sequences was 64 000 000, which was 21-times higher than 10 years ago. During this time the number of protein structures in the Protein Data Bank increased only 3-fold and nowadays has reached 110,000. To decrease the disparity between primary and tertiary protein structures, protein structure modelling methods are being developed. Protein structure reconstruction from a contact map generates collection of models containing properly oriented and mirror models. This is because all of them share the same contact map. Properly oriented protein models and mirror models can constitute competitive forms in nature. Our main goal is to identify the indicators which could be useful in distinction the mirror models without a priori knowledge about the structure. We assumed that some of the PyRosetta energy terms will be significantly different for mirror models than for properly oriented models.

In our work we used protein models which were reconstructed from contact maps of experimental SCOP domains with our tool – C2S_pipeline. We investigated 100 models for each of 1305 domains. With Biopython we calculated structural features of the models and with PyRosetta we computed the energy terms, whose linear combination is the total energy of the model.

C2S_pipilne generates mirror models and properly oriented models with the same probability of 0.5. The structural quality of the properly oriented models and mirror models is comparable. In all-alpha domains the energy term which describes electrostatic energy (*hack_elec*) offered the most reliable indicator between properly oriented and mirror models (for 77% domains). Simultaneously, the energy terms related to the probability of amino acid at dihedral angles Ψ and Φ (*p_aa_pp*) and with the Ramachandran preferences (*rama*) were statistically different for 68% and 64% domains. Despite the intuition that the mirror images of the protein riches in alpha-helices are easier to identify, we observed more energy terms which were significantly different for more than 75% domains with beta-sheets. These energy terms were also *rama* and *p_aa_p*, additionally the attractive and repulsive portions of the Lennard-Jones potential (*fa_atr*, *fa_rep*) and Lazaridis-Karplus solvation energy (*fa_sol*).

P20.10

Fusion of information acquired from gene expression and semantic similarity measure of genes using GO database in order to reduce gene signature

Wojciech Labaj, Andrzej Polanski[1]

Institute of Informatics, Silesian University of Technology, Akademicka 16, Gliwice, Poland
Wojciech Labaj <wojciech.labaj@polsl.pl>

DNA microarray is one of the modern high-throughput experiment analysis techniques for gene expression, which can measure in parallel hundreds of thousands of targets in the form of genes or their alternative variants. There are routines for analyzing results of the experiment including statistical testing with false discovery corrections. They result in gene signatures: lists of genes used for summarizing all of the analysis steps.

Despite the existence of established routines, there are still challenges in the field of appropriate selection of gene signature, forasmuch as there are problems in instability of composition, defining size (number of genes) of gene signature, which can lead to unreliability of results of biological inference. Therefore there are many efforts towards improving algorithms for construction of gene signature.

Moreover, one of commonly used databases for biological inference, which gathers information about genes, is Gene Ontology database. It concentrates widely available knowledge from biological experiments and scientific publications. On this basis a particular gene is annotated to special set of GO terms, which describes its membership in terms of the ontology: Biological Process, Molecular Function or Cellular Component. The structure of GO database allows to measure semantic similarity between GO terms and also, what is important, between genes. Our idea is to take into account accumulated knowledge and use it in an alternative approach.

In this paper we would like to introduce an idea of reduction of gene signature and show preliminary results of the analysis. We used semantic similarity measure to calculate the average similarity of each gene to the remaining genes in gene signature and then merge this information with information extracted from gene expression analysis. Our step can help to select genes, which are specific for analyzed case. This can improve further classification steps or select genes, which can enhance biological interpretation of the experiment.

We compare different methodologies of reduction of gene signatures for part of MILE experimental data. The comparison is based on classification quality as well as measuring of gene signature stability.

Acknowledgements:

This work was financially supported by SUT grants BKM/515/RAU-2/2015.

P20.11

Functional genomic data analysis of irradiation impact to ATF3 protein expression

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

[1]Silesian University of Technology, Institute of Automatic Control, Data Mining Group, Akademicka 16, 44-100 Gliwice; [2]Public Health England, Chilton, Didcot, OX11 ORQ, United Kingdom; [3]King Faisal Specialist Hospital & Research Centre, Riyadh 11211, Kingdom of Saudi Arabia
Agata Szymanek <agatszy050@student.polsl.pl>

Aim: Aim of the study was to identify single nucleotide polymorphisms (SNP) related to changes in activating transcription factor 3 (ATF3) expression in response to ionizing radiation. Functional analysis of genes containing candidate SNPs was performed to identify possible pathways related to ATF3 expression changes in exposed cells.

Materials and methods: Investigated sample was containing 44 unrelated healthy Caucasian individuals, from whom T lymphocytes were collected. Analysed data set was containing genotyping results of 567,095 SNPs, and ATF3 qPCR expression measured in normal conditions (0Gy) and after irradiation with treatment dose of 2Gy. Fold Change (FC) of signals was specified to obtain the information about ATF3 expression change. For each SNP three models of genotype-phenotype interactions were considered: genotype, dominant and recessive and appropriate statistical test were performed. Afterwards, Benjamini-Hochberg (BH) correction for multiple testing was performed. Genomic locations of candidate SNPs were identified and signalling pathways investigation was performed using Fisher exact test to determine their significance in radiosensitivity.

Results: Group of p-values were obtained for each SNP. The best model was chosen by lowest p-value criterion, and with significance level α=0.05 the 69,267 polymorphisms were obtained. After BH correction for multiple testing, number of significant SNPs decrease to 226. Fisher exact test was performed to find signalling pathways overrepresented within candidate genes. The analysis revealed 20 overrepresented dis-regulations, where relation to melatonin signalling pathway seems to be the most significant to phenomenon of radiosensitivity. For the following pathway SNPs in genes PLCB1 and PRKACB were found (for both 2 SNPs in intron region) and the structure of signal cascades to ATF3 was detected.

Conclusions: Benjamini-Hochberg correction for multiple testing provides statistically reliable results. Still, more steps are necessary to validate biological significance. In our studies, relation between ATF3 and melatonin signalling was identified. Statistical analysis is important part of biological studies and should be introduction to biological validation.

Acknowledgement:

This work was funded by HARMONIA 4 no. 2013/08/M/ST6/00924 (JP), SUT grant BKM/506/Rau1/2016/t.26 (JZ), and 11-BIO1429-20 for (GA). Calculations were carried out using infrastructure of GeCONiI (POIG.02.03.01-24-099/13). The functional analyses were generated through the use of QIAGEN's Ingenuity Pathway Analysis.

P20.12

Mixture Modeling of 2D Gel Electrophoresis Images Enhances Quality of Spot Detection

Michal Marczyk

Data Mining group, Institute of Automatic Control, Silesian University of Technology, Poland
Michal Marczyk <Michal.Marczyk@polsl.pl>

2D gel electrophoresis is the most important and widely used method for analysis of complex protein mixtures. The use of this technique has been effectively defined in many cases to disclose both physiological mechanisms and proteins associated with clinical pathologies that can aid in the discovery of biomarkers. The lack of efficient, effective, reproducible and reliable methods for 2D gel analysis has been a major factor limiting the contribution of 2DGE to biomedical research on a wider scale.

In this paper we check if the quality of detection of protein spots estimated by existing software can be improved by use of the mixture of 2D Gaussian functions. Spots detected by existing software are used as initial conditions in estimating mixture model parameters. To perform a comprehensive comparison study, we created tens of artificial 2DGE gel images. We assumed that the observed image is an effect of accumulating background, random noises and real information about protein abundance. The information about distributions of all noises and spots was estimated using real dataset.

Comparison based on a large number of artificially generated datasets proved that in case where there are clusters of overlapping spots, like in scenarios with the highest number of true spots, the mixture model enables detecting components hidden behind others. Model components are better characterized by the accurate spot position and shape than spots detected by existing software. Fitting the mixture model to synthetic image allows for achieving higher sensitivity in detecting spots and better overall performance of the spot detection. These results show a great potential of using mixture model in gel image analysis.

P20.13

Functionality status of murine tcr sequences – comparative analysis of diversity indices and effect size statistics

Justyna Mika[1], Serge M. Candéias[2], Christophe M. Badie[3], Joanna Polanska[1]

[1]Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science, Gliwice, Poland; [2]French Alternative Energies and Atomic Energy Commission, Laboratory of Chemistry and Biology of Metals, Grenoble, France; [3]Public Health England, Cancer Mechanisms and Biomarkers group, Radiation Effects Department, Didcot, United Kingdom
Justyna Mika <justyna.kotas@polsl.pl>

T lymphocytes are one of the cells responsible for adaptive immune response. Each T cell carries at its surface a receptor (TCR) whose sequence, complementary to antigenic peptides, is created during VDJ recombination process. It consists in the selection and assembly of one of each variable (V), diversity (D) and joining (J) gene segments into an exon coding for complementarity determining regions of TCRs. The pool of TCR sequences might be divided into two groups regarding its functionality status, productive or nonproductive, depending on presence of stop codons or the absence of open reading frames between V and D gene segments.

The following study was performed on data coming from high-throughput sequencing of murine whole blood TCRs. Samples were collected from 30 mice, control or exposed to 0.1Gy and 1Gy irradiation at three time-points (1, 3 and 6 months post irradiation). Globally, more than 0.5 million unique sequences were obtained with count differences between individual samples reaching 12.5 fold change between two extreme samples. To strengthen the power, pooled analysis was performed with relation to proper dose and time-point. To compare the functionality status of sequences modified Hutcheson t-test was proposed using Pielou's diversity index. Variance for each sample was calculated through bootstrap sampling with drawing numbers from multinomial distribution (with 2,000 iterations). A Bonferroni correction for multiple testing was applied. To strengthen the reliability of results, for each comparison an effect size was calculated using modified Cohen's d distance.

Every sample was carrying more productive than nonproductive sequences. The performed analysis resulted in nine Pielou's indices with confidence intervals indicating significant differences between each sample. The diversity of sequence status changes with time into less homogenous, favoring productive sequences over nonproductive ones. Different response might be observed for irradiation with small (0.1Gy) and intermediate (1Gy) doses, with effect stable in time. Size of the observed effect is rather small due to big sample sizes, however consistent in individual mice.

Acknowledgements:

This work was financially supported by the European Commission (DoReMi, European Atomic Energy Community's 7th FP 2007-2011, grant agreement no.249689), and NCN grant Harmonia 4 register number 2013/08/M/ST6/00924. Calculations were carried out using GeCONil infrastructure (POIG.02.03.01-24-099/13)

P20.14

An integrative approach vs restrictive thresholds for combining gene expression data sets on radiation response

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

[1]Data Mining Group, Institute of Automatic Control, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland; [2]Institute of Computer Science, Silesian University of Technology, ul. Akademicka 16, 44-100 Gliwice, Poland
Anna Papież <anna.papiez@polsl.pl>

Combining information from high-throughput cellular biology data sets is becoming an essential tool for scientific researchers. The never ceasing growth of data available through repositories implies the urge of raising the processing algorithms' efficiency, as large amounts of meaningful information are being omitted in this deluge of experimental results.
In this work, we demonstrate the utility of statistical combination techniques for merging two expression data sets for a radiation study of breast cancer patients in order to increase the yield of biologically relevant conclusions.
The research was carried out on two independent data sets obtained in the course of microarray experiments. In both settings lymphocyte RNA from breast cancer blood donors was extracted and subject to a high dose of ionizing radiation. The data required careful normalization and accounting for batch effects in order to enable comparative analysis due to the difference in experimental platforms: olignucleotide and cDNA.
Merging the two data sets was performed using two approaches: restrictive - taking into account the intersection of differentially expressed genes assigned with a fixed threshold in both experiments and integrative - based on the integration of p-values from both experiments.
The algorithms resulted in candidate irradiation signature gene lists. These lists were assessed by determining the data set separability with regard to dose. Furthermore, they were then used to build Leave-One-Out classifiers by means of logistic regression and the performance was evaluated with Positive and Negative Predictive Values.
The integrative algorithm proves to be efficient for improving the findings in gene expression signature research. Merging of test statistics enables a more meticulous incorporation of the differentiating strength of a particular gene feature, rather than setting a fixed p-value threshold. This provides a quantitative and qualitative increase in the quantity of gene signatures. These findings may ensure a more comprehensive material for inference by specialists from the biological and medical studies.

Acknowledgements:

This work was supported by SUT grant BKM/506/RAU1/2016/t.30 (AP), NCN Harmonia grant register number DEC-2013/08/M/ST6/00924 (JP). Calculations were carried out using GeCONiI infrastructure funded by project number POIG.02.03.01-24-099/13.

P20.15

Microbiological Resources and Bioinformatic Tools at MultiGenBank Database

Krzysztof Pawlik[1], Paweł Mackiewicz[2], Magdalena Kotowska[1], Anna Tobiasz[1], Klaudia Kozub[1], Agnieszka Korzeniowska-Kowal[1]

[1]Institute of Immunology and Experimental Therapy, Polish Academy of Sciences, 12 Rudolf Weigl St., Wrocław, 53-114, Poland; [2]Department of Genomics, Faculty of Biotechnology, University of Wroclaw, 14a Fryderyk Joliot-Curie, 50-383 Wrocław, Poland
Krzysztof Pawlik <pawlik@iitd.pan.wroc.pl>

MultiGenBank is an bioinformatics project carried outat the Institute of Immunology and Experimental Therapy, funded under action 2.3: Investments in development of science infrastructure, IEOP 2007-2013. The aim of this project is to create an online platform containing two database modules. The first one contains human genetic polymorphisms data associated with diseases, whereas the second module is designed as a microbial database including genetic data of microorganisms harbouring potentially useful genes.
The microbiological module contains records corresponding to microorganism strains, to a great extent derived from the Polish Collection of Microorganisms (PCM). For particular strains, the database gathers information about their genotypes, taxonomy, sequences of main molecular markers (16S rRNA, ITS-1 and ITS-2), additional markers (e.g. dnaJ, sod A, rpoB, gyrB, recA, groEL, dnaK) and results of genotyping by RFLP, RAPD or other methods. Data, both newly submitted and present in the database, obtained by molecular techniques based on differences in lengths of DNA fragments (e.g. RFLP and RAPD) can be compared and analysed. This function is dedicated for comparison of microorganisms on the strain or isolate level. The database can be searched for homologous sequences using BLAST algorithm. The sequences found can be aligned using Clustal Omega and simple phylogenetic trees may be constructed.
The microbial module is associated with a synthetic biology tool -SynKeD- which is useful in designing new synthetic polyketides and gene sequences encoding their synthesis pathways. SynKeD implements the idea of using well-known and described modules of modular type I PKS, which can be considered as “bricks” for construction of new polyketide synthesis pathways. The microbial module also includes nucleotide sequences of modules from known PKS assigned for particular carboxylic acids units, which are incorporated into a polyketide chain. Based on this data, SynKeD responds to the query containing a chemical formula of a polyketide chain. As a result, the application returns a aminacids sequence of PKS modules which can be used for the synthesis of the desired polyketide chain.

Acknowledgements:

Project was co-financed by the European Union through the European Regional Development Fund under the Operational Programme Innovative Economy. Grants for innovation. We invest in your future.

P20.16

Tabu search algorithm for RNA Partial Degradation Problem (RNA PDP)

Agnieszka Rybarczyk[1,2], Alain Hertz[3], Marta Kasprzak[1,2], Jacek Błażewicz[1,2]

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland; [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland; [3]Department of Mathematics and Industrial Engineering, Ecole Polytechnique and GERAD, Montreal, Canada;

Agnieszka Rybarczyk <arybarczyk@cs.put.poznan.pl>

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

In this work, we would like to present our recent results concerning the RNA degradation process. In our studies we used artificial RNA molecules designed according to the rules of degradation developed by Kierzek and co-workers [1, 2]. On the basis of the results of their degradation, we have proposed the formulation of the RNA Partial Degradation Problem (RNA PDP) and we have shown that the problem is strongly NP-complete [2]. We would like to propose a new efficient heuristic approach, in which two tabu search algorithms cooperate. The algorithm can reconstruct a given RNA molecule, having as input the results of the biochemical analysis

of its degradation, which possibly contain errors (false negatives or false positives). Results of the computational experiment, which prove the quality and usefulness of the proposed method, are presented.

References:

1. Kierzek R (2001) *Methods Enzymol* **341:** 657-675.
2. Blazewicz J *et al* (2011) *Journal of Computational Biology* **18:** 821-834.

P20.17

EMQIT: a web tool for energy based PWM matrices quality improvement

Karolina Smolińska[1,2], Marcin Pacholczyk[1], Marek Kimmel[1,3]

[1]Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland; [2]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland; [3]Department of Statistics, Rice University, Houston, TX, USA

Karolina Smolińska <karolina.smolinska@polsl.pl>

Transcription factor binding sites (TFBSs) are important in gene regulation. Developing new methods of modelling of TFBSs structures in DNA is very important. Traditional models are based on Position Weight Matrices (PWMs) obtained either computationally or from experimental data.

In previous work, we presented a modification of the approach introduced by Alamanova *et al.* [2]. We observed that tuning of Boltzmann factor weights, used for conversion of calculated energies to nucleotide probabilities, has a significant impact on the quality of the associated PWM matrix (Smolinska *et al.*, 2016, unpublished). Consequently, we developed EMQIT (Energy Matrices Quality Improvement Tool), a web server that use tuned Boltzmann weights and ROC curves to obtain better predictive models of transcription factor (TF) binding sites. Presented tool is written in R/Shiny package [2]. EMQIT require PWM matrix and name of TF to perform quality improvement calculations. The resulting PWM is displayed in the main page, as a PWM matrix and logo. Moreover, EMQIT compares improved PWM with matrix available in TRANSFAC [3] and matrix created by implementation of original Alamanova et al. method (3DTF server [4]). Model summary is available as logos and matrices. To test our tool we used, energy matrices generated by 3DTF server [4] as an input. We applied our method to data available for p50p50, p50p65, p50RelB, p53, HSF1 and Erα.

The comparison has shown a significant similarity and comparable performance between the calculated and the experimental matrices (TRANSFAC). Improved 3DTF matrices achieved significantly higher AUC values than the original 3DTF matrices (at least by 0.05) and, at the same time, detected notably more experimentally verified TFBSs. The presented approach can be successfully used to any energy based PWM matrix.

References:

1. Alamanova D, Stegmaier P, Kel A (2010) *BMC Bioinformatics* 11.
2. Rstudio, Inc, Easy web applications in R, 2013, http://shiny.rstudio.com/.
3. Matys, Vea *et al* (2003) TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31.1**: 374-378.
4. Gabdoulline R, Eckweiler D, Kel A, Stegmaier P (2012) 3DTF: a web server for predicting transcription factor PWMs using 3D structure-based energy calculations. *Nucleic Acids Res* **40**: W180-W185.

Acknowledgements:

This work has been supported by Polish National Science Centre funds under grants OPUS DEC-2012/05/B/NZ2/01618 and BKM-514/RAU1/2015 p.9 based at the Institute of Automatic Control, Silesian University of Technology.

P20.18

New in silico approach to assess RNA secondary structures with non-canonical base pairs

Natalia Szostak[1,3], Agnieszka Rybarczyk[1,2,3], Maciej Antczak[1,3], Tomasz Zok[1,3], Mariusz Popenda[2,3], Ryszard Adamiak[1,2,3], Jacek Blazewicz[1,2,3], Marta Szachniuk[1,2,3]

[1]Institute of Computing Science, Poznan University of Technology, Poznań, Poland; [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland; [3]European Center for Bioinformatics and Genomics, Poznan University of Technology, Poznań, Poland
Natalia Szostak <nszostak@cs.put.poznan.pl>

RNA function depends on its structure, therefore an appropriate recognition of the latter is of great importance. One particular concern is the assessment of base-base interactions, described as the secondary structure. It greatly facilitates an interpretation of RNA function and allows for structure analysis on the tertiary level. The RNA secondary structure can be predicted from sequence using in silico methods often adjusted with experimental data, or assessed from 3D structure atom coordinates. Computational approaches consider mostly Watson-Crick and wobble base pairs. Handling of non-canonical interactions, important for a full description of RNA structure, is still a challenge.
Here we present novel two-step in silico approach to asses RNA secondary structures with non-canonical base pairs. Its idea is based on predicting the RNA 3D structure from sequence or secondary structure that describes canonical base pairs only, and next, back-calculating the extended secondary structure from atom coordinates. We have integrate in a computational pipeline the functionality of two fully automated, high fidelity methods: RNAComposer for the 3D RNA structure prediction and RNApdbee for base pair annotation. We have benchmarked our pipeline on 2559 RNAs sequences with the size up to 500 nucleotides obtaining better accuracy in non-canonical base pair assessment than the compared methods that directly predict RNA secondary structure.

P20.19

A Petri net based method for comparison of biological processes

Bartłomiej Szawulak[1], Piotr Formanowicz[1,2]

[1]Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 60-965 Poznań, Poland, [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland
Bartłomiej Szawulak <bszawulak@cs.put.poznan.pl>

Petri nets are a mathematical formalism used to model and analyze concurrent processes, especially in technical sciences. However, in recent years there is a growing interest in an application of nets of this type to model complex biological systems. On one hand, Petri nets have an intuitive graphical representation, and on the other hand, there are many strict mathematical methods for analysis of their properties. Moreover, they are very well-suited for describing in a clear way a structure of a modeled system. This makes them an interesting alternative for differential equations. However, biological systems have their own specificity, so new methods for analysis of their Petri net based models are still being developed. One of the important and still not solved problems in this area is a comparison of such models. An effective algorithm solving this problem would allow to determine common structures occurring in different, possibly related, biological systems, what may have a great impact on understanding their properties.
In this work we propose such a method which is based on an analysis of subnets and relations between them. In our approach subnets are defined by ADT sets. A similarity of the compared Petri nets is calculated on the basis of a match between nodes (i.e. places and transitions). The method has been tested on models of various metabolic networks and preliminary results of these test are promising.

P20.20

The obtaining recombinant *Gallus gallus* YGP40 and protein analysis by bioinformatics tools

Agnieszka Szmyt, Anna Dąbrowska, Józefa Chrzanowska

Department of Animal Products Technology and Quality Management, Wroclaw University of Environmental and Life Science, Chelmońskiego 37/41, 51- 630 Wrocław, Poland
Agnieszka Szmyt <a.tatomir.szmyt@gmail.com>

The glycoprotein YGP40, which is released by cathepsin Dfrom the C-terminal fragment of chicken vitellogenin-2 (VTG2), is the source of several peptides with immunomodulatory activity. Those polypeptides complex, named yolkin, possess biological properties similar to mammalian colostrinin. It has an ability to stimulate human whole blood cells to release the pro-inflammatory interleukin IL-6 and anti-inflammatory IL-10. Furthermore, these peptides complex is able to stimulate neuronal cells to secrete the mature form of brain-derived neurotrophic factor (BDNF), which, together with IL-6, plays an important role in the control of central nervous system functionality. It was also reported, that yolkin reveals an anti-inflammatory activity, as it stimulates IL-10 releasing, which level decreases during neurodegenerative processes. Yolkin also moderates the aging symptoms, supports learning functions. Recent data has shown significant influence of yoklin on behavior and cognitive functions of Wistar rats (young and old). It seems, that yolkin complex may also have positive therapeutic effect in human neurodegenerative diseases, like Alzheimer disease.

The natural source of yolkin is the laying hen's egg yolk. The applied method allowed to isolate heterogenous group of yolk-derived peptides, alongside IgY purification. Nevertheless, the purification yield is low and time- consuming, as it takes about 4–5 days to obtain maximum of 32 μg of yolkin from one egg yolk. In our studies we focused on heterologous expression of YGP40 gene in *Escherichia coli* BL21 as a host.

In the first step of studies, we designed the gene of interest sequence containing specific restriction sites, basing on back-translation of VTG2 amino acids sequence. The plasmid DNA material obtained from host cells, after previous gene expression, was sequenced and analyzed with BLAST program. The obtained protein sequence was analyzed *in silico* for the presence of various potential bioactive peptides. The physical parameters of YGP40 and secondary structure prediction was determined, as also the modeling the course of hydrolysis through selected proteolytic enzymes.

P20.21

Bioinformatic analysis of motifs in microRNA vicinity in plants

Joanna A. Kowalska[1], Katarzyna Tomczyk[1], Joanna Sarzyńska[2], Marta Szachniuk[1,2]

[1]Institute of Computing Science, Poznan University of Technology, Poznań, Poland, [2]Institute of Bioorganic Chemistry PAS, Poznań, Poland
Katarzyna Tomczyk <katarzyna.rybicka@cs.put.poznan.pl>

Recently, many scientists have become interested in microRNA (miRNA) – small non-coding RNA molecules which participate in a regulation of gene expression on post-transcriptional level. In animals, these particles are responsible for the regulation of cellular processes and they are associated with diseases, like cancer and neurodegenerative diseases. In plants, miRNA plays a key role in the response to stress conditions, such as sudden changes of temperature, drought or nutrient deficiency, and it participates in the process of growth and development.

Biogenesis of plant miRNA differs from that in animals. Hairpin structure of animal pre-miRNA is cleaved to the form of miRNA-miRNA* duplex after being transported out of the nucleus. In contrast, plant miRNA is cleaved before its export to the cytoplasm. The other important difference lies in enzymes mediating the miRNA maturation process. Dicer, animal RNase III enzyme, serves as 'molecular ruler' in animals, while Dicer Like 1 (DCL1) is responsible for cutting out miRNA-miRNA* duplex in plants. While the mechanism of mature microRNA production is almost fully understood for animals, the biogenesis of miRNA in plants remains unclear. This research aims to support understanding the principle of enzyme DCL1 by searching for structural patterns that could lead DCL1 to the cleavage sites.

Here, we present pattern searching routine and the results of its application for precursor structures of Arabidopsis thaliana family. By using available bioinformatics tools, i.a. WebLOGO, RNAstructure, MCQ4Structures, RNAComposer, and own scripts, we try to identify structural patterns in the vicinity of miRNA. We present potential motifs found in sequences and secondary structures. We also show the first results of tertiary structure analysis based on predicted 3D models of miRNA precursors, validated with respect to its alignment with the DCL1 model.

P20.22

RNAComposer: new developments to increase accuracy of 3D RNA structure prediction

Tomasz Zok[1,3], Maciej Antczak[1,3], Mariusz Popenda[2,3], Joanna Sarzynska[2,3], Tomasz Ratajczak[1,3], Ryszard W. Adamiak[1,2,3], Marta Szachniuk[1,2,3]

[1]Institute of Computing Science, Poznan University of Technology, Poznań, Poland; [2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland; [3]European Center for Bioinformatics and Genomics, Poznan University of Technology, Poznań, Poland
Tomasz Zok <tomasz.zok@cs.put.poznan.pl>

The tertiary structure of RNA determines its functions in the cellular processes. Unfortunately, among vast number of RNA molecules known through their nucleotide sequence, only a small fraction has their atomic coordinates recognized. This shortage of complete structural information is partially rooted in the difficulty of 3D structure determination process especially associated with complex RNAs. Therefore, *in silico* prediction of RNA 3D models is very important for the development of modern structural biology. Around a dozen different approaches has been already proposed and used for this problem (Dufour *et al*, 2015). One of them is RNAComposer developed by our group – a fully automatic system able to model large RNA 3D structures (Popenda *et al*., 2012). The method was found useful for many predictions and has gained recognition and popularity in the community. However, some target molecules are still a challenge to model and there is a room for improvements in our method (Miao *et al*., 2015). Recentlym, we have introduced new developments in the RNAComposer webserver to give users more control over the 3D structure modelling process. Three new *in silico* tools to predict secondary structure from sequence have been incorporated and a customised way of providing 3D elements into the pipeline has been added. The introduced functionality significantly improves the accuracy of prediction. This is presented by an in-depth analysis of 3D models obtained for precursors of miR160 family members.

References:

Dufour D *et al* (2015) *Wiley Interdisciplinary Reviews: Computational Molecular Science* **5**: 56-61.
Popenda M *et al* (2012) *Nucleic Acids Research* **40(13)**: e112.
Miao Z *et al* (2015) *RNA* **21**: 1-19.

Acknowledgements:

This work was supported by grants from National Science Center, Poland [2012/05/B/ST6/03026, 2012/06/A/ST6/00384].