Autumn Workshop PTBI 2020

Polish Bioinformatics Society

November 24, 2020

Contents

Schedule

26 November 2020	
Welcome	9:00
Session 1	9:10
Session 1 Q&A	9:40
Break	9:50
Session 2	10:00
Session 2 Q&A	10:30
Break	10:40
Session 3	10:50
Session 3 Q&A	11:30
Break	11:40
Session 4	11:50
Session 4 Q&A	12:20
Break	12:30
Session 5	12:40
Session 5 Q&A	13:20
End of day 1	13:30

27 November 2020	
Welcome	9:00
Session 6	9:10
Session 6 Q&A	9:50
Break	10:00
Session 7	10:10
Session 7 Q&A	11:00
Break	11:10
Best MSc-s	11:20
Best MSc-s Q&A	11:50
Break	12:00
Best PhD-s	12:10
Best PhD-s Q&A	12:40
End of day 2, awards, and closing remarks	12:50

Abstracts

In silico exploration of quadruplex structures

26 Nov
9:10

Joanna Miśkiewicz[1], Mariusz Popenda[2], Joanna Sarzyńska[2], Tomasz Żok[1, 3], Marta Szachniuk[1, 2]

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

G-quadruplexes are unique structures present in nucleic acids. They are located in G-rich sequences in eukaryotic, prokaryotic, and viral genomes. G-quadruplexes, also called G4s, play an important regulatory role in biological processes, including regulation of gene expression. G4's basic unit is a G-tetrad, a composition of four guanines arranged in pseudo-planar position. To build a G-quadruplex, at least two G-tetrads are needed. The conformations of G4s drew the medical community's attention to potential applications in drug development strategies against cancerous and neurodegenerative diseases. Currently, many research laboratories explore and study quadruplexes using experimental and computational methods.

Here we present our in silico research on quadruplexes. We started investigating quadruplexes focusing on their secondary structure topology and proposed a new classification named ONZ. Along with the ONZ classification, we also enhanced dot-bracket notation by including a 2-line format, to represent quadruplexes. With the newly developed automated method, ElTetrado, we categorized all tetrads and quadruplexes with reference to ONZ classes. All quadruplex structures found in PDB repository are categorized by ONZ classes and will be stored in our quadruplex database – ONQUADRO. We also investigated known G4-tools for quadruplex prediction and visualization, and we explored quadruplex databases. The result of this investigation is a complex G4 survey which describes 35 bioinformatic resources that concern quadruplexes. We believe that our quadruplex research will help scientists across different laboratories to better understand the nature of G4s.

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

26 Nov
9:20

RNA junctions from a 3D structure perspective

Jakub Wiedeman[1], Maciej Antczak[1, 2], Jacek Kaczor[1], Maciej Miłostan[1, 3], Tomasz Żok[1], Marta Szachniuk[1, 2]

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

The determination of 3D structures of biomolecules is one of the biggest challenges in the bioinformatics field. Over the last few years, we could observe a growing number of new methods and initiatives around predicting three-dimensional structures of proteins and RNAs. While prediction methods successfully determine quite many structure elements, some of them still need adjustments. Multibranched loops are examples of structural motifs that are hard to predict accurately by most of the computational approaches. In our work, we created the RNAloops database that collects all multibranched loop structures found in RNAs. The motifs are regularly updated and identified using a self-developed search algorithm operating on dot-bracket representations of the input structures. The presented database contains the proposition of a novel representation for outgoing loops. It utilizes Euler angles to determine their spatial relation which is enriched by information, i.a. sequence or secondary structure. Our analysis of gathered data showed that we can find structures that contain over 100 different n-way junctions and for the analyzed set one in eight structures had at least one junction. Therefore, we strongly believe that n-way junctions have a huge impact on overall RNAs folds, and obtained data can be used in the process of modeling unknown RNA 3D structures and refinement of the existing ones.

This project has been supported by the National Science Centre, Poland
[2016/23/B/ST6/03931], [2019/35/B/ST6/03074], and the statutory funds of Poznan University of Technology, Poland.

The influence of structure size on similarity metrics values.

26 Nov
9:30

Piotr Kłosowicz[1], Tomasz Żok[2]

[1] Adam Mickiewicz University of Poznan

[2] Poznan University of Technology

A comparison of tertiary structure is very important in order to understand RNA function. It can be used to identify certain motifs in newly-found molecules, which helps to understand their function without conducting too many experiments. It also applies to structure alignment, clustering, or designing new molecules. But in order to make a valid comparison, there is a necessity for the reliable measure. In order to achieve that, a group of scientists from the University of North Carolina in their article "On the significance of an RNA tertiary structure prediction", written in 2010, determined the distribution of root-mean-square deviation (RMSD) for fourteen RNA native structures and a set of decoys. Their work has shown that RMSD value is highly dependent on the size of the molecule and in order to compromise that, they suggested using a prediction significance (P-value), which helps to evaluate, if the given prediction is better than expected by chance for the molecule of that size.

But there are more measures that can be applied to compare tertiary structures of the RNA than just RMSD. In our research, we are concentrating on the mean of circular quantities - MCQ, which calculates the difference between molecules based on the torsion angles between nucleotides. We want to inspect, if it follows the same problem with scaling with size, that RMSD has, and, if necessary, propose a value, that will help to make more reliable comparisons. We are going to use in our studies the same molecules that were used in research on RMSD and P-value, and then compare their real tertiary structure with those generated by suitable software.

26 Nov
10:00

RNAlign2D- RNA sequence and structure multiple alignment tool, based on pseudo-amino acids substitution matrix

Tomasz Woźniak[1], Małgorzata Sajek[2], Jadwiga Jaruzelska[1], Marcin Sajek[1]

[1] Institute of Human Genetics, Polish Academy of Sciences, Poznań, Poland
[2] Department of Human Molecular Genetics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznań, Poland

The function of RNA molecules is mainly determined by their secondary structure. Addressing that issue requires creation of appropriate bioinformatic tools that enable alignment of multiple RNA molecules to determine functional domains and/or classify RNA families. The existing tools for RNA multiple alignment that use structural information are relatively slow. We developed an extremely fast Python based RNAlign2D tool. It converts RNA sequence and structure to pseudo-amino acid sequence and uses customizable scoring matrix to align RNA secondary structures and sequences using MUSCLE. It is suitable for RNAs containing modified nucleosides and/or pseudoknots. Our approach is compatible with virtually all protein aligners.

Mining biomacromolecular interactions with the BioShell package

26 Nov
10:10

Justyna Kryś, Monika Pikuzinska, Dominik Gront

Faculty of Chemistry, University of Warsaw

Functioning of every living cell starts with an interaction between two biomolecules. The knowledge of how exactly molecules interact is very important in understanding virtually any biological process. Molecular modelling is a technique commonly used in deciphering life on a molecular level. Outputs from these simulations usually comprise 3D structures and their energies, evaluated in a given force field. These tools however rarely explain how the biomolecules interact.

In the newest version of the BioShell software package we provide an application that calculates not only close-distance contacts but also can recognise an interaction type such as: hydrogen bond, Van der Waals and stacking. Detection of these interactions requires an all-atom input structure that also includes hydrogen atoms. Such models however are not always available. Therefore we also introduced an algorithm to reconstruct missing atoms in a protein structure. The newest BioShell 3.1. update also expands the package with unit tests framework and includes an HSSP format reader. During the presentation I will provide detailed information about these new BioShell features.

26 Nov
10:20

fingeRNAt - a novel tool for high-throughput analysis of nucleic acid-ligand interactions

Natalia Szulc[1, 2], Zuzanna Mackiewicz[1], Janusz Bujnicki[1, 3], Filip Stefaniak[1]

[1] Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland

[2] Laboratory of Protein Metabolism in Development and Aging, International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland (current affiliation)

[3] Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland

Nucleic acids are becoming increasingly attractive targets for drugs. Since most targets of small molecule drugs are proteins, the portfolio of nucleic acids-oriented bioinformatics tools is limited. Here we present fingeRNAt - a novel and open-source software for calculation of Structural Interactions Fingerprints (SIFs) for nucleic acid - ligand complexes. SIFs translate information about 3D interactions into a binary string, where the respective bit in the fingerprint is set to 1 in case of detecting particular interaction, and to 0 otherwise. By using SIFs, the interactions are represented in a unified fashion, thus allowing for easy analysis and comparison, as they provide a full picture of all interactions within the complex.

Our program detects non-covalent interactions in nucleic acid - ligand complexes and directly converts them to SIFs. The inputs are (i) structure of RNA or DNA and (ii) structures of its ligands, which can be small molecules, proteins or nucleic acids. In our implementation SIFs can be calculated at low, medium or high resolution allowing the user to choose the desired details level, thus making our program applicable to many problems. The calculated output, containing SIF for each nucleic acid-ligand complex, is easy to process and can be directly visualized by our program. We will present the design of the fingeRNAt program and discuss the application of generated fingerprints to various bioinformatic analyses such as analysis of the occurrence of non-covalent interactions in RNA-ligand complexes or comparison of binding modes in complexes of nucleic acid with different ligands.

Binary genome maps assembly

26 Nov
10:50

Przemysław Stawczyk, Robert Nowak
Institute of Computer Science, Warsaw University of Technology

Reduction in cost of sequencing genomes provided by next-generation sequencing technologies greatly increased number of genomic projects. As a result, there is growing need for better methods for validate assemblies, as well as better methods for scaffolding. One of the promising ideas is to use heterogenious data in assembly project. Optical Mapping (OM) was found very useful in validating genomic assemblies, correction and scafffolding, single raw OM read describe long fragment of DNA molecule.

Raw OM data are could by asembled to create consensus maps. The result of such assembly are maps that can cover entire chromosome. Process of assembly is computationally hard because of a large fraction of errors in input data. There are very few algorithms for creating whole genome restriction maps other than proprietary software from sequencer manufactures. Available projects was based upon dynamic programming with backtracking which scales very poorly with large genomes. Moreover these algorithms require existence of high quality reference genome, they could be used only in resequencing projects.

In this work we depict new algorithm and computer program to assembly OM reads without reference genome. In our algorithm we explore possibility for using binary representation for genome maps. We focused on the efficiency of data structures and algorithms, as well as the ability to scale on parallel platforms.

In our in silico experiments we obtained whole genome of *e.coli* genome using 10x coverage and *BspQI* enzyme. For *Caenorhabditis Elegans* using *BspQI* enzyme and x25 coverage our simulations returns 9 contigs covering 6 chromosomes of reference genome.

26 Nov
11:00

A new overlap graph method for DNA sequence assembly

Sylwester Swat[1], Artur Laskowski[1], Jan Badura[1], Wojciech Frohmberg[1], Paweł Wojciechowski[1, 2], Aleksandra Świercz[1, 2], Marta Kasprzak[1], Jacek Błażewicz[1, 2]

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

Reconstruction de novo of a genome sequence is a great challenge, largely due to computational difficulties connected with processing millions of reads at once. ALGA is a new method realizing this process and is based on the overlap-layout-consensus approach. The approach consists of three phases: construction of the overlap graph, preparation of the graph for traversal and agreement of final sequences. It is generally viewed as more accurate than the so-called de Bruijn graph

approach, but much more consuming in the sense of time and memory. Several new ideas were implemented in order to increase efficiency at each of the phases. ALGA offers options enabling assembly based on data of different quality. It was tested on several real data sets, and the results were evaluated with the standard tool QUAST. In comparison to other assemblers, ALGA provides very good results according to metrics such as genome coverage fraction, length of resulting sequences and occurences of misassemblies.

Automated approach for sequence profile generation

26 Nov
11:10

Marcin Piwowar[1, 2], Dominik Gront[1]

[1] Faculty of Chemistry, Univeristy of Warsaw

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

Despite the recent progress in the field, construction of a multiple sequence alignment (MSA) still requires a considerable effort from a human expert. Automated methods can make various errors, that often result from an unfortunate selection of input sequences, e.g. when set of these sequences is redundant. In this contribution I used tools from BioShell package to filter an input sequence set and construct better MSA in an iterative fashion. At every stage of the algorithm overrepresentation was reduced and the worst sequences which decreased the quality of a resulting MSA were removed. The method has been tested and validated on a non-redundant set of sequences from HOMSTRAD database.

26 Nov
11:20

In silico evaluation of SARS-CoV-2 primers performance

Michał Kowalski[1], Alina Frolova[1, 2], Witold Wydmański[1], Wojciech Branicki[1]
Paweł Łabaj[1]

[1] Małoposka Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[2] Institute of Molecular Biology and Genetics of National Academy of Sciences of Ukraine, Kyiv, Ukraine

Emergence of novel coronavirus SARS-CoV-2 had became a global threat in a blink of an eye. Many research groups, corporations and world organisations had proposed specific primers sets for RT-qPCR as for primary diagnostic tests to evade further spreading of virus by being able to isolate infected people until their will be able to rejoin society. Over time, studies and reports had shown that not every proposed set of primers can amplify the virus thus generation of falsly negative results became a serious problem. Since global lockdown hadn't been handled properly in many countries, evolution of SARS-CoV-2 had became region specific, which introduced mutations that altered performance of globally recommended primers.

Our group inspired by diagnostic work of our collegues from Human Genome Variation Research Group from Malopolska Centre of Biotechnology and from MetaSUB consortium had developed a Python library specifically for task of benchmarking primer sets along with probes. Currently analysis of recommended primers performance had been evaluated on set of over 90k highest quality sequences from GISAID database and 223 sequences from Poland (GISAID database mixed with sequences from Human Genome Variation Research Group). Analysis showed that this threat shouldn't be generalized to global standards and every region should have a set of primers designed for rapid detection of novel coronavirus as a part of localy based prevention system. We conclude that meta information aware software for design of primers and probes is in urgent need and we pursue the goal of delivering such software following open source philosophy.

Estimated nucleotide reconstruction quality symbols of basecalling tools for Oxford Nanopore sequencing

26 Nov
11:50

Wiktor Kuśmirek

Warsaw University of Technology, Institute of Computer Science

Currently, one of the fastest growing DNA sequencing technologies is nanopore sequencing. One of the key stages of processing sequencer data is the basecalling process, which from the input sequence of currents measured on the pores of the sequencer reproduces the DNA sequences called DNA reads. Many of the applications dedicated to basecelling together with the DNA sequence provide the estimated quality of reconstruction of a given nucleotide. Herein, we examinated the estimated quality of nucleotide reconstruction reported by another basecallers. The results showed that the estimated reconstruction quality reported by different basecallers may vary depending on the tool used. In particular, for some tools, along with successive symbols of the estimated reconstruction quality (which theoretically should mean more and more accurate reconstruction), the real quality of the nucleotide increases (the number of matched nucleotides increaces and the number of errors decreases). However, there are tools that report the estimated reconstruction quality in the basecalling results, but these values are in no way interpretable. What is more, the estimated reconstruction quality reported in basecalling process is not used in any investigated tool for processing nanopore DNA reads.

26 Nov
12:00

Dig (deep)er - Deep learning algorithms for the imbalanced classification of correct and incorrect SNP genotypes from WGS pipelines

Krzysztof Kotlarz[1], Magda Mielczarek[1, 2], Tomasz Suchocki[1, 2], Bartosz Czech[1], Bernt Guldbrandtsen[3], Joanna Szyda[1, 2]

[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2] Institute of Animal Breeding, Balice, Poland

[3] Animal Breeding Group, Department of Animal Sciences, University of Bonn, Bonn, Germany

A downside of NGS technology is high technical error-rate. The aim of applications of Deep Learning algorithms is classifying NGS-based SNPs into the correct and the incorrect calls. The dataset consisted of whole genome DNA sequences (WGS) of four Danish Red Dairy Cattle bulls, sequenced by the Illumina platform. Additionally, the bulls were typed using the Illumina BovineHD BeadArray. The first step was SNP classification into correct (i.e. a SNP genotype from WGS was concordant with a SNP genotype from a microarray) or incorrect otherwise. The training data set was composed of data from three bulls and comprised 2,274,915 SNPs among which 2% were incorrect. The validation data set consisted of data from the fourth bull with 749,506 correct (98.05%) and 14,940 incorrect SNPs. With such an imbalance in class counts, the traditional classification methods based on logistic regression fail. Therefore, a deep learning algorithm implemented in Keras was used to build a classifier for SNPs, based on standard sequence quality statistics available from the variant calling output. Three deep learning algorithms were applied: (i) a baseline algorithm, (ii) an algorithm with class weights and (iii) an algorithm that oversamples the class of incorrect SNPs. For each of the algorithms, in addition to a standard threshold (0.5), probability threshold values were also estimated based on the optimisation of F1 or SUMSS metrics. The results showed that the most parsimonious baseline algorithm and an algorithm with the weighting of SNP classes provided the best classification of the validation data set. Both classified 19% of truly incorrect SNPs as incorrect and 99% of truly correct SNPs as correct and resulted in the F1 score of 0.21 – the highest among the compared algorithms.

DNA sequence features underlying large-scale duplications and deletions in human

26 Nov
12:10

Mateusz Kołomański[1], Joanna Szyda[1, 2], Magdalena Frąszczak[1], Magda Mielczarek[1, 2]

[1] Biostatistics group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

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

The 1000 Genomes Project characterized over 88 million of genetic variants, including single nucleotide polymorphisms, indels, transposons and Copy Number Variants (CNVs). The latter are large structural variants manifested mainly as duplications and deletions. They may cover up to 12% of the genome and have impact on phenotypic diversity and disease. The aim of the study was to characterize regions of human genome that are susceptible to CNVs. Our study used 5867 structural duplications and 33181 structural deletions available from the 1000 Genomes Project. Regions covered by CNVs and 100 bp-long sequences flanking those regions were extracted from reference genome. The sequences were considered in the context of unknown nucleotides, Guanine-Cytosine pairs, sequence complexity and functional annotation. Four sets of randomised regions were extracted from reference genome for comparison.

The number of unknown nucleotides in the analysed dataset was relatively low, which confirmed high reliability of studied variants. The distribution of fraction of GC pairs within CNV regions differed significantly from distribution in randomized sets of sequences. The distribution of content of low-complexity sequences in duplications was not significantly different from the randomized data, but in deletions it was. 100-bp regions located downstream and upstream of duplications, as well as downstream of deletions were significantly different from the randomised data, but sequences located upstream of deletions were not different. The majority of variants intersected with gene regions - mainly with introns.

In conclusion, the most interesting finding was a significant interplay between low complexity regions and CNVs as well as impact of CG pairs content on CNV formation.

26 Nov
12:40

Canonical Correlation- based bioinformatic analysis for effective melanoma biomarker discovery.

Sonia Wróbel[1], Ewa Stępień[1], Cezary Turek[2], Monika Piwowar[2]

[1] Department of Medical Physics, Jagiellonian University, Marian Smoluchowski Institute of Physics, Krakow, Poland

[2] Department of Bioinformatics and Telemedicine, Jagiellonian University–Medical College, Krakow, Poland

Next Generation Sequencing (NGS) and other advanced large-scale experimental methods provide enormous amounts of multi-dimensional biological data. Understanding the interactions between transcriptomics, proteomics and other types of data generated using different platforms is fundamental. In such analyzes, the integration of multiple OMICS datasets together and selection of variables is the key to obtaining interpretable results. Canonical Coronation Analysis (CCA) is one of the most powerful methods for this bioinformatic challenge. Over the last years, a number of promising results for implementing CCA in the integration of OMICS data have been proposed. In order to be able to analyze multi-OMICS datasets in the context of systems biology, an optimal approach to data integration, analysis and interpretation should be developed.

Here we introduce a new method based on canonical correlation analysis that uses real life dataset to meet the challenge of cancer biomarker discovery. The bioinformatics pipeline was successfully applied to human skin melanoma multi-OMICS dataset containing: (1) microvesicle-micro-RNA transcriptomics, (2) microvesicle proteomics, (3) cell-total-RNA transcriptomics. The method applies a sparse CCA (sCCA) to three matrices, starting from features correlation across integrated experimental data. Validation using clinical data as well as supporting meta-data from databases allows the identification of evidence-based candidates for highly significant molecular signatures like melanoma-associated microRNAs and oncoproteins.

REST and ZBTB33 in glioma REST ChIP-seq peaks - partners or competitors?

26 Nov
12:50

Marta Jardanowska[1], Bartosz Wojtas[2], Malgorzata Perycz[1, 2], Bozena Kaminska[2], Michal J. Dabrowski[1]

[1] Institute of Computer Science, Polish Academy of Sciences, Poland
[2] Nencki Institute of Experimental Biology, Warsaw, Poland

The canonical role of REST transcription factor (TF) is regulation of neurogenesis and glial cells development. Therefore, REST is described as both activator and repressor of transcription, depending on physiological or pathophysiological context.

The purpose of this study was to check whether other TF motifs are in close proximity to REST transcription factor binding sites (ChIP-seq) from U87 cell line. To each peak we assigned its summit and within the 200bp around the summit, we searched for the other TF motifs. Based on TCGA glioma RNA-seq and inhouse REST ChIP-seq data, it was specified whether REST represses or activates the expression of the particular gene, based on the correlation results, negative or positive, respectively.

With this approach, we identified TF motifs for the REST activated and repressed genes. Strikingly, in the top of the motifs ranking for the REST activated genes were KAISO motifs, characteristic for the ZBTB33 transcription factor. KAISO motifs had higher frequency and lower q-value in the REST activated genes while in REST repressed genes REST motifs performed better. Analysis of the nucleotide sequences of the REST and KAISO motifs showed that they differ significantly, meaning that the co-occurrence of these TF motifs within the examined sequences was not due to the sequence similarity. We also observed that in the REST activated genes, KAISO motifs were more frequent in the close proximity to the REST-peak summits.

These results, by showing the co-occurrence of two TF motifs within the short sequences derived from REST ChIP-seq, and putting them in the context of gene activation or repression, suggests that while the main REST role is repressive, its role within the promoters of activated genes might be co-dependent on ZBTB33.

26 Nov
13:00

DNA methylation patterns of active enhancers specific for *pilocytic astrocytoma* and Higher Grade Glioma samples.

Agata Dziedzic[1], Marta Jardanowska[1], Marcin Grynberg[2], Michał J Dąbrowski[1]

[1] Institute of Computer Science, Polish Academy of Sciences

[2] Institute of Biochemistry and Biophysics, Polish Academy of Sciences

Gliomas are the most common tumors of the central nervous system, and are among the most deadly types of cancer. There is evidence that epigenetic changes contribute to modulating gene expression in glioma and therefore are considered to be one of the main drivers of glioma development. In this study we compared gene regulation mechanisms via enhancers in *pilocytic astrocytoma* (PA) to Higher Grade Glioma (HGG) samples (*diffuse astrocytoma* and *glioblastoma* histopathological grades). Locations of active enhancers were determined by peaks of H3K27ac in non-promoter regions. Gene expression and methylation levels of single cytosines were obtained from RNA-seq and Bisulfite-seq experiments, respectively.

We analysed 114 enhancers specifically active in PA and 124 in HGG samples. HGG-specific enhancers had lower frequency of guanine and cytosine nucleotides (46% GC content) then PA-specific enhancers (54% GC content) and higher global DNA methylation level - frequency of hyper to hypo-methylated was higher in HGG samples. Within enhancer sequences there were 230 TF motifs identified. Methylation pattern of 14 motifs was confirmed to be consequently hypermethylated in HGG compared to PA samples (chi-square test). At least one gene targeted by each of the enhancers with specific motifs, differed in expression between PA and HGG samples. Some of these genes are known to contribute to gliomagenesis.

These results indicate specific TF motifs whose methylation may have an influence on regulation of TG expression and therefore contribute to gliomagenesis.

Novel alternative splicing events detection in human genome with Spladder

26 Nov
13:10

Agata Muszyńska[1, 2], Paweł P. Łabaj[1, 3]

[1] Małopolska Centre of Biotechnology UJ, Krakow, Poland

[2] Institute of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland

[3] Chair of Bioinformatics, Boku University, Vienna, Austria

Although human genome is widely studied since many years its complexity remains not fully understood. One of the mechanisms that stands for that is alternative splicing, which is a process of joining exons in multiple ways, so that novel mRNA and, in fact, novel proteins are produced. Currently we are not fully aware of all of the splicing events that might be present in a given genome. One of the tools that provides the possibility to investigate that is Spladder. It builds an augmented splicing graph, based on current annotation and then expands it with novel events. Currently Spladder supports detecting six different types of such events. We used Spladder software on data from SEQC consortium project.

We investigated 3 samples (A- mixture of 10 different cancer cell lines, B- healthy individual and C- A and B samples mixed in 1:1 ratio) run on different RNA targeting panels, as well as on whole transcriptome sequencing data obtained with two protocols- ribo-depletion and polyA selection. Preliminary results show that there is a fraction of genes containing novel events, which seems to be cancer or sample specific, but majority is the same irrespective of sample. It seems that the current gene model can be extended by this data. Spladder also revealed that the fraction of intron retention events is higher for whole transcriptome sequencing data than for targeted approach and is higher for ribo-depletion protocol than for polyA selection, what is expected after comparing sample processing and library preparation for these approaches.

These results show that there is still a lot of work ahead of us to fully describe our genome but at the same time that Spladder might be a good tool, not only for that task, but also for others like detecting cancer specific events.

27 Nov
9:10

BioMetaNet: Meta-Network model for human lymphoblastoid cell lines representing complete biological interactome

Kaustav Sengupta[1, 2, 3], Michal Denkiewicz[1, 4], Anup Kumar Halder[5], Subhadip Basu[5], Dariusz Plewczynski[1, 3]

[1] Center of New Technologies, University of Warsaw, Warsaw
[2] Faculty of Mathematics Informatics and Mechanics, University of Warsaw
[3] Doctoral School of Exact and Natural Sciences, University of Warsaw, Warsaw
[4] Faculty of Mathematics and Information Science, Warsaw Technical University, Warsaw
[5] Department of Computer Science and Engineering, Jadavpur University, India

Most biological processes take place due to the interactions of various biomolecules and therefore, can be represented as graphs. We propose a meta-network representation of the complete map of DNA pairwise interactions for human lymphoblastoid cell lines together with the proteins and metabolic pathways. We integrate into the single graph multiple biological networks, namely, Chromatin Interaction Network (CIN), Genomic Association Network (GAN), Protein-Protein Interaction Networks (PIN), Gene Ontology (GO) terms, and pathways.

In our approach, the whole genome 3D structure is represented as a graph with nodes matching the anchors from ChIA-PET experiments and CTCF chromatin loops defining the edges. The GAN is then built by mapping genes to the nearest anchors, and further extended by linking genes with the corresponding proteins mapping GAN onto an undirected PIN. The GO terms are integrated with two interacting proteins as a representation of a given interaction functionality. Finally, the nodes of PIN representing proteins are linked to the nodes of the metabolic functional pathway graph. The meta-network model exhibits the graph centrality properties different from that of the random models. Furthermore, using functional analysis of SNPs from GWAS, we identified one of the DNA region (chr6:32014923-33217929) characterized by the statistically significant enrichment of SNPs related to autoimmune diseases. We analyzed the meta-network and found two proteins present in the specific location share a higher concentration of pathways between them. These shared pathways can be critical for studies of autoimmune diseases. Therefore our meta-network model can be instrumental in getting a complete picture of biological functionality linked with 3D chromatin interactions.

MASTERMIND: The Best Linear Model to Accurately Determine Monoisotopic Mass

27 Nov
9:20

Piotr Radziński, Michał Startek, Anna Gambin

Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland

Nowadays, monoisotopic mass is used to be an important feature in top-down proteomics. Knowing the exact monoisotopic mass enables precise and quick protein identification in large protein databases. However, only in spectra of small molecules monoisotopic peak is visible, for bigger molecules position of the peak have to be predicted. By improving prediction of the peak, we contribute to more accurate identification of molecules, what is crucial in fields such as chemistry and medicine. In this work we present MASTERMIND algorithm, that is a two-step procedure to predict monoisotopic mass for proteins with 8-400 kDa mass range. The first step is to approximate monoisotopic mass by linear regression based on average mass and variance of a given spectrum. The second step rounds linear prediction to the closest point which is reliable to be a peak in the spectrum. For 96.6% of proteins, prediction error is below 0.2 ppm, what is approx. 30% better than in recently proposed MIND tool. Our algorithm was implemented in python, data analysis was performed in R. Proteins to learn the model comes from Uniprot database, their theoretical spectra were calculated by use of IsoSpec structure calculator.

27 Nov
9:30

Novel approach to search for interdigitated proteins - unusual domain swapped topology

Mateusz Skłodowski[1], Joanna M. Macnar[1, 2], Dominik Gront[1]

[1] Faculty of Chemistry, University of Warsaw

[2] Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw

Protein domain swapping is a well-known phenomenon, studied by numerous researchers for many years [1]. In the specific case presented by this contribution, beta strands are swapped in a way that a beta sheet is formed by pieces from two alternating chains. A path taken across a sheet, i.e. along hydrogen bonds that connects its strands, leads through chains A, B, A, B and so on. Hence the name of such unusual protein topology: interdigitated structure. The seminal work describing an interdigitated protein has been published more than a decade ago[2]. In this work, we for the first time present the results of research for interdigitated proteins present in the whole PDB. The analysis we have undertaken was to identify the proteins that have the motive described earlier. For this purpose, the BioShell package [3], [4] and graph theory were used, and then polypeptides were grouped according to the amount of beta strands present in the beta-sheets. For further analysis, a group of proteins with the longest six-element beta sheet was adopted, in which their structural, sequential and functional similarity was studied. The obtained results indicate that interdigitated proteins are typically small homodimers.

[1] Liu, Y., & Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. Prot Sci, 11(6), 1285-1299

[2] Wang, S, et al. The crystal structure of the AF2331 protein from Archaeoglobus fulgidus DSM 4304 forms an unusual interdigitated dimer with a new type of $\alpha + \beta fold.ProtSci.18.11(2009) : 2410 - 2419$

[3] Gront, D. & Kolinski, A. (2006). BioShell—a package of tools for structural biology computations. Bioinformatics, 22(5), 621-622

[4] Macnar, JM. et al. BioShell 3.0: Library for Processing Structural Biology Data. Biomolecules 10.3 (2020): 461

Analysis of small molecules parameters in ligand-protein complexes

27 Nov
9:40

Joanna M. Macnar[1, 2], Wladek Minor[3], Dominik Gront[1]

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

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

[3] Department of Molecular Physiology and Biological Physics, University of Virginia, Charlottesville, USA

Structural information of ligand-macromolecule complexes is critical for biomedical sciences, especially for structure-based drug discovery and structural bioinformatics. The majority of experimental information about macromolecules complexed with ligands is coming from X-ray crystallography. The combination of structural, biochemical, spectroscopic, and bioinformatics methods may revolutionize drug discovery, albeit only when the structural information is very accurate.

The availability of synchrotron sources, with modern equipment and software, has led to the availability of roughly 160,000 structures deposited in the Protein Data Bank (PDB). Unfortunately, a small number of crystal structures deposited in the PDB are of suboptimal quality, including some with incorrectly identified and modeled ligands in protein-ligand complexes. The BioShell 3.0 package is a modern structural bioinformatics toolkit that includes the analysis of ligand-protein complexes. Combining a graph-theoretical approach, kernel density estimators, bioinformatics methods, and chemical knowledge, we present an analysis which checks the correctness of selected ligands from PDB deposits. This analysis will lead to an improved library of restraint parameters and subsequently better refinement of ligand-protein complexes.

27 Nov
10:10

Cost-sensitive feature selection - information theory approach

Tomasz Klonecki

Institute of Computer Science Polish Academy of Sciences

Feature selection is a crucial problem in many bioinformatics tasks. Usually, the considered variables are cheap to collect and store but in some situations, the acquisition of feature values can be problematic. For example, when predicting the occurrence of the disease we may consider the results of some diagnostic tests which can be very expensive. The existing feature selection methods usually ignore costs associated with the considered features. The goal of cost- sensitive feature selection is to select a subset of features that allow to predict the target variable (e.g. occurrence of the diseases) successfully within the assumed budget.

The main purpose of this research is to review filter methods of feature selection based on information theory and to propose new variants of these methods taking into account feature costs. We try to solve NP-hard feature selection problem using various greedy algorithms.

We support our theoretical reasoning by experiments on artificial data as well as on the large clinical database MIMIC-III. The experiments confirm that introduced methods can improve the performance of disease detection algorithms with a constrained budget.

Hierarchical clustering in search for the most relevant variables in small-n-large-p datasets

27 Nov
10:20

Radosław Piliszek, Witold Rudnicki

Computational Centre, University of Bialystok

Gene expression and genomic datasets from biomedical studies belong to the so-called small-n-large-p class. Such datasets describe a relatively small number of objects (records) - counted in tens, hundreds and thousands - using a large number of variables (features) - counted in tens, hundreds and thousands of thousands.

Many machine learning algorithms suffer performance decreases in such a case. Moreover, human analysis of the studied phenomenon is severely hampered. Various feature selection algorithms have been proposed to tackle this problem. However, there might still exist many relevant features. A naive approach of top-N ranking will usually discard relevant information and still keep sets of variables carrying the exact same information. Eliminating correlations upfront is of no use because correlation does not map exactly to information about the decision variable.

We propose an approach to limit the number of variables further by clustering variables using an existing measure of relevant variable discovery and scoring - the MDFS - MultiDimensional Feature Selection. We searched for clusters of variables having relatively negligible information gain between themselves. Each cluster is then replaced by the cluster representative variable. There are, however, several ways to build such clusters, even when constrained to hierarchical methods. There are also different ways to choose the representative.

We present the preliminary results of several variants of this approach, based on an analysis of datasets from the CAMDA 2017 challenge on Neuroblastoma-diagnosed patients. The patients in the three datasets are described using respectively CNV (Copy Number Variation), GE (Gene Expression) profiles from microarrays and GE profiles from RNA-seq.

27 Nov
10:30

Exploring the microbiome protein structure space using simulations and deep learning

Paweł Szczerbiak[1], Douglas Renfrew[2], Julia Koehler Leman[2], Daniel Berenberg[2], Chris Chandler[2], Vladimir Gligorijevic[2], Richard Bonneau[2], Tomasz Kościółek[1]

[1] Małopolska Centre of Biotechnology, Jagiellonian University, Kraków, Poland
[2] Flatiron Institute, Simons Foundation, New York, USA

Human microbiome comprises about 3 million unique bacterial genes which is 100 more than the number of human body genes. Exploring them would give us a possibility to treat diseases that originate in or are influenced by the human microbiome. Main goal of the Microbiome Immunity Project [1] is to understand the role played by the various bacteria in the human microbiome. In the first stage of the project we want to map all proteins produced by those bacteria. For this purpose we prepared a dataset consisting of 300,000 unique newly predicted microbial protein structures (in future we want to reach 1,000,000). In order to make the analysis more robust, we used two methods: Rosetta [2] and DMPFold [3] which utilize different approaches to the protein structure prediction problem. In the poster we are showing the difference between both methods with special emphasis on new folds identification and structure space visualization. We also plan to create an open access database that anyone can use in their own analysis.

[1] https://www.worldcommunitygrid.org/research/mip1/overview.do
[2] https://www.sciencedirect.com/science/article/pii/S0076687904830040
[3] https://www.nature.com/articles/s41467-019-11994-0

Comprehensive functional annotation of metagenomes and microbial genomes using deep learning-based methods

27 Nov
10:40

Mary Maranga[1], Paweł P. Łabaj[1], Richard Bonneau[2], Tommi Vatanen[3, 4], Tomasz Kościółek[1]

[1] Małopolska centre of biotechnology, Jagiellonian university, Poland
[2] Flatiron Institute, New York, NY, USA
[3] Liggins Institute, University of Auckland, New Zealand
[4] Broad Institute, Cambridge, MA, USA

The human gut harbors a numerous number of microbial species. Recent evidence suggests that the intestinal microbes contributes to the development and persistence of diseases such as type-1 diabetes (T1D), ulcerative colitis and obesity. However, the exact mechanisms of how gut microbiota influences health remains poorly understood. This study aims to characterize the function potential of the human gut microbiome in type-1 diabetes.

We used DIABIMMUNE infant human gut microbiome data (1067 samples) as a case study. This data was previously collected in Finland, Estonia and Russian Kareli . We developed a custom metagenomics annotation pipeline based on DeepFRI , a machine learning protein function prediction method that combines LSTM and deep learning Graph Convolutional Networks (GCN) to functional profile metagenome gene functions. Our method integrates de novo genome reconstruction, taxonomic profiling and functional annotation.

We generated a non-redundant gene catalog comprising of 1.9M genes. We used a genome quality threshold of >90% genome completeness and <5% contamination, the final genomes matching these criteria were 2256. Using DeepFRI function predictions we observed an increase in annotation coverage compared to EggNOG predictions. We annotated approximately 74% with DeepFRI and 20% with EggNOG. Work is under-way to improve our functional annotations using structure based DeepFRI method. We provide a computational methodology for annotation of human gut metagenome sequences and connecting its functionalities with human health. The findings of this work will provide information which could be explored to design better therapeutic strategies to prevent immune mediated disorders.

27 Nov
10:50

Standardizing 16S rRNA gene sequencing downstream analysis for Oxford Nanopore and Ion Torrent technologies

Katarzyna Kopera[1], Dedan Githae[1], Maria Kulecka[2], Jerzy Ostrowski[2], Paweł P. Łabaj[1], Tomasz Kościółek[1]

[1] Malopolska Centre of Biotechnology, Jagiellonian University, Cracow, Poland
[2] Department of Gastroenterology, Hepatology and Clinical Oncology, Centre of Postgraduate Medical Education, Warsaw, Poland

16S rRNA marker gene sequencing is a staple technique for microbiome analyses that provides rapid and cheap bacterial identification. The most popular and well-standardized experimental technique is based on Illumina short-read sequencing. Alternative techniques are long-read Oxford Nanopore (ONT) and short-read IonTorrent platform (PGM). While both producers provide complete 16S analysis workflows, they are often not fully transparent, unadaptable, and limited to the basic methodology implemented within a given workflow. This produces a community-wide need for more in-depth workflows which at the same time will validate the applicability of the two sequencing methods in the area of 16S experiments.

We describe the powers and limitations of the two methods (PGM and ONT) by comparing them with our alternative downstream analysis created in QIIME2. The workflow was tested on 16S data generated on the Oxford Nanopore's and Thermo Fisher's sequencing machines and their 16S metagenomics kits. Data from 126 fecal samples from mice humanized with human stool were analysed. Different diversity metrics, taxonomy classification, and differential abundance methods were performed. For 21 common samples, Mentel test and Procrustes were made to compare the correlation of beta diversity between the two platforms.

We have managed to achieve powerful and more refined results using the approach we created, despite the limitation of information imposed by manufacturers' policies. Mentel test and Procrustes suggest good correspondence of the results from the two platforms. However, we would like to stress the further need for the entire community to cross-validate results and develop new standardized approaches for the data produced from PGM and ONT 16s sequencing solutions.

Best MSc Thesis Competition laureates

Patterns of DNA variation between the autosomes, the X chromosome and the Y chromosome in Bos taurus genome

27 Nov
11:20

Bartosz Czech[1], Bernt Guldbrandtsen[2, 3], Joanna Szyda[1, 4]

[1] Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[2] Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University, Tjele, Denmark

[3] Department of Animal Sciences, University of Bonn, Bonn, Germany

[4] Institute of Animal Breeding, Balice, Poland

The new $ARS-UCD1.2\_Btau5.0.1Y$ assembly of the bovine genome has considerable improvements. That might be assumed that a more accurate identification of patterns of genetic variation can be achieved with it. The aim of this study was to explore differences in genetic variation between autosomes, the X chromosome, and the Y chromosome. In particular, single-nucleotide polymorphisms (SNPs), densities of insertions and deletions (InDels), annotation, InDel lengths, nucleotide divergence, and Tajima's D were compared between chromosomes. Whole-genome DNA sequences of 217 individuals representing different cattle breeds were examined. The analysis included the alignment to the new reference genome and variant calling. 23,655,295 SNPs and 3,758,781 InDels were detected. In contrast to autosomes, both sex chromosomes had negative values of Tajima's D and lower nucleotide divergence. That implies a correlation between nucleotide diversity and recombination rate, which is obviously reduced for sex chromosomes. Moreover accumulation of nonsynonymous mutations on the Y chromosome could be associated with loss of recombination. Also, the relatively lower effective population size for sex chromosomes leads to a lower expected density of variants.

27 Nov
11:30

Multiple Sequence Alignment Analysis

Paulina Dziadkiewicz, Norbert Dojer

Faculty of Mathematics and Information Science, Warsaw University of Technology

The constant growth of biological data intensifies the development of algorithms for genomic analysis. One of the goals of modern bioinformatics is the effective processing of genomes originating from different organisms. This implies the need to create their common representation – called pan-genome. Solutions for this problem should be universal – to be able to handle multiple data formats, effective – to process sequences of any length and functional - to enable analysing the genomic sequences jointly or use it as a reference.

This work was focused on providing a tool for discovering and visualizing the relationships between sequences. The pan-genome data model is built from a multiple sequence alignment as a partial order graph. It is then used to perform taxonomic analysis and create a hierarchic division of aligned sequences called Affinity Tree. Each node of the Affinity Tree has assigned a subset of genomes, as well as their homogeneity level and averaged consensus sequence. Moreover, subsets assigned to sibling nodes form a partition of the genomes assigned to their parent.

The second part of the presented solution is visualisation software. It presents results of the algorithmic part in a browser app. This visualisation enhances analysis of both – pan-genome and Affinity Tree.

The algorithm was applied to two types of MSA datasets: simulated (yielded from genome sequence evolution simulations) and real-life (computed for Ebolavirus genomes). The resulting Affinity Trees are compatible with corresponding phylogenetic trees so it can serve as both taxonomic study and a population reference pan-genome.

Despite the fact, that the software does not exhaust the pan-genome subject, it successfully shows the alignment of multiple sequences as a hierarchical data structure.

Cerebral Microbleedsdetectionon MR imageswith hybrid neural network

27 Nov
11:40

Aleksandra Suwalska[1], Yingzhe Wang[2], Ziyu Yuan[3], Yanfeng Jiang[3, 4], Jinhua Chen[5], Mei Cui[2], Xingdong Chen[3, 4], Chen Suo[3, 6], Joanna Polanska[1]

[1] Silesian University of Technology, Department of Data Science and Engineering, 44-100 Gliwice, Poland

[2] Department of Neurology, Huashan Hospital, Fudan University, Shanghai, People's Republic of China

[3] Fudan University Taizhou Institute of Health Sciences, Taizhou, People's Republic of China

[4] State Key Laboratory of Genetic Engineering and Collaborative Innovation Centre for Genetic and Development, School of Life Sciences, Fudan University, Shanghai, People's Republic of China

[5] Taizhou People's Hospital, Taizhou, People's Republic of China

[6] Department of Epidemiology Ministry of Education Key Laboratory of Public Health Safety, School of Public Health, Fudan University, Shanghai, People's Republic of China

Cerebral Microbleeds (CMBs) are very important imaging marker in the diagnosis of Cerebral Small Vessel Diseases. Manual inspection of CMBs is time-consumingand prone to human errorsbut the existing automated or semi-automated methods have insufficient detection accuracy. Data sets with CMB cases arelimited, therefore, the analysis is complicated and many proposed solutions require more than one MRI modality, which are not always available.

In the study, weproposed an effective system for the automated detectionof Cerebral Microbleeds based on a hybrid neural network which uses both numeric and image data as inputs.We trained and tested the system on 304 patients with 39CMBs in total (Dataset 1). The developed system wasalso tested on an independent dataset (Dataset 2) with61 patients and 21 CMBs in total.Theproposed approach requires one MRI modalityonly and consists of five main steps: (1) images' pre-processingand standardization, (2) detection of thecerebral bleeds, (3) filtrationof the potentially non-CMB objects, (4) calculation of the predefined shape and texture featuresfor every candidate CMB, and (5) object classificationby the hybrid NN.The values of the obtained quality indices are very promising. We achieveda sensitivity of 90.0% and a specificity of 98.5% forDataset 1 and a sensitivity of 91.5% and specificity of 95.2% forDataset 2.

Despitethe small number of CMB cases, thestudyconfirms the usefulness of deep learningcombined with mathematical modelling. The implemented two-input mixed-data neural network has never been used before and reached the lowest number of false positives per patient (0.54) compared with other existing methods(from 107.5 to 1.6).The results also show that the combination of numeric and image data provides sufficient knowledge for the detection of Cerebral Microbleeds.

FundingThis work was funded by the National Key Research and Development Program of China (2017YFC0907000 [to Dr. Chen and Dr. Suo], 2017YFC0907500 [to Dr. Chen],2016YFC0901000 [to Dr. Chen]), the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01 [to Dr. Chen]), the China Post-doctoral Science Foundation (2019M661376 [to Dr. Jiang]),the Key Technology Research and Development Program of Taizhou (TS201833 [to Dr. Yuan]),the

Key Research and Development Plans of Jiangsu Province, China (BE2016726 [to Dr. Chen]), and the International Science and Technology Cooperation Program of China (2014DFA32830[to Dr. Chen]) .This work wasalsofunded by the National Science Centre, Poland, grant 2015/19/B/ST6/01736 [to Prof. Polanska]. Additionally, MSc Suwalskais a holder of European Union scholarship through the European Social Fund (grant POWR.03.05.00-00-Z305).

Best PhD Thesis Competitions laureates

Modeling the structure and dynamics of proteins using coarse-grained models with different resolution scale

27 Nov
12:10

Aleksandra Badaczewska

Faculty of Chemistry, University of Warsaw, Warsaw, Poland

My PhD dissertation describes a low-resolution coarse-grained SURPASS model and its applications in molecular modeling of protein structure and dynamics. In particular, a unique deeply simplified representation of the protein structure was proposed and statistical potentials of a knowledge-based force field were derived. The concept of SURPASS representation is very simple and assumes averaging of short secondary structure fragments. The specific interaction model distinguishes the protein-like SURPASS chain from a random polymer. Statistical potentials describe local structural regularities characteristic for most globular proteins. SURPASS model was used for replica exchange Monte Carlo dynamics simulation of single-domain proteins, with secondary structure as the only sequence-dependent input data for the interaction model. Despite its deep simplification, SURPASS model reproduces reasonably well the basic structural properties of proteins. The model allows very efficient sampling of the entire conformational space of protein. SURPASS model is a tool that overcomes the limitations of coarse-grained moderate resolution models, which are still too expensive to model efficiently large biomolecular systems and their interactions. SURPASS model can be easily adapted as an initial step to various multiscale modeling strategies.

27 Nov
12:20

Algorithms and models for protein structure analysis

Aleksandra I. Jarmolińska[1, 2], Joanna I. Sułkowska[1], Anna Gambin[2]

[1] Centre of New Technologies, University of Warsaw

[2] Faculty of Mathematics, Informatics and Mechanics, University of Warsaw

The last decade has seen a large increase in the number of studies related to protein topology. Currently, there are over 1500 known knotted or slipknotted protein chains (Dabrowski-Tumanski et al., 2019), and almost 10 000 protein links (Dabrowski-Tumanski et al., 2016). Screening of available RNA structures has also found entanglements (Micheletti et al., 2015). Recent advances in the study of chromatin structure gave rise to new 3D models—many of which contain entanglements, including composite knots. Still, the subject of molecular entanglements remains relatively unknown to a lot of researchers, including those studying protein structures. One obvious reason is the steep learning curve for actually seeing the knots in a 3D structure visualization. Knot_pull allows an easy analysis of topological intricacies by providing the user with a trajectory of smoothing steps—from the full structure, to the minimal number of coordinates preserving the original topology (with regard to fixed position of chain termini) — and the knot type (including separation of composite knots, and indication of any linking present) - without using the prevalent probabilistic approach.

Studying the sequences of entangled proteins also encounters problems - finding the most closely related protein family may require detecting the similarity based on sequence profiles, which are not easily (multiple-)aligned. To overcome this obstacle, I introduce two new heuristic for creating a multiple profile alignment, by using a modified Dijkstra's shortest path tree algorithm to find the maximum weight trace (Kececioglu, 1993) of a set of pairwise alignments. This allows for an easy, large scale comparison of loosely related protein groups.

Integrative data analysis methods in multi-omics molecular biology studiesfor disease of affluence biomarker research

27 Nov
12:30

Anna Papież, Joanna Polańska

Department of Data Science and Engineering, Silesian University of Technology, Gliwice, Poland

The need for transforming large amounts of data in the life sciences drives the development of statistical and data mining algorithms for merging and validation of biomedical experiments. Although this issue has been previously commonly acknowledged in the scientific community, the constantly increasing amounts of data require continuous efforts towards the optimization of data analysis pipelines. Therefore, the aim of this thesis is to investigate diverse approaches for high-throughput molecular biology integrative data analysis to enable the discovery of disease of affluence biomarkers.

The work consists of a detailed overview of existing advancements in high-throughput molecular biology techniques data integration,followed by the demonstration of novel algorithms for combined analysis of data derived from multi-platform and multi-domain experiments. Initially, an original batch effect identification algorithm based on dynamic programming is presented, as correcting for these effects constitutes a part of the intra-experiment data integration pipeline. Its performance on identifying batch structure is proven to be highly efficient, and moreover, batch effect pre-processing entails potential new knowledge discovery in studied diseases and conditions.Subsequently, two microarray data sets obtained using different platforms for biomarker research in breast cancer patients are analyzed to highlight the potential of measurement transformation to achieve computational and biological consistency. The statistical and data-mining integrative approaches with functional validation and profile modeling provides a comprehensive solution for elucidating dose response mechanisms and potential biomarker signatures. Moreover, custom statistical integrative methods applied to a transcriptomics and proteomics data set on ischemic heart disease plutonium mine workers enabled discrimination of dose dependent protein expression changes from the age dependent changes and validation of pathways identified previously in the proteomics data. Another approach to data integration, which enabled the identification of factors playing a key role in differentiation of irradiated samples, was conducted on multi-tissue exosome proteomics data.