ENBIK2018 conference proceedings

Editors

Petr Čech, Daniel Svozil

Prague 2018

Organizing committee

Assoc. Prof. Daniel Svozil, Ph.D.
Ing. Petr Čech, Ph.D.

Scientific committee

Assoc. Prof. Daniel Svozil, Ph.D.
Mgr. Michal Kolář, Ph.D.
RNDr. Petr Bartůněk, CSc.
Prof. Jan Holub, Ph.D.
RNDr. Jiří Vondrášek, CSc.

ENBIK2018 conference proceedings

Printed by powerprint s. r. o.
Brandejsovo nám. 1219/1, 165 00 Praha 6 – Suchdol

Published by the University of Chemistry and Technology, Prague
Technická 5, 166 28 Praha 6, Czech Republic

ISBN 978-80-7592-017-1

SESSION 1

Tools

L1-01

fuzzyreg: An R package for fuzzy linear regression

Martínková N.[1], Škrabánek P.[2]

[1] *The Czech Academy of Sciences, Institute of Vertebrate Biology, Brno, Czech Republic*
[2] *Department of Process Control, Faculty of Electrical Engineering and Informatics, University of Pardubice, Pardubice, Czech Republic*

Fuzzy linear regression provides a possibilistic alternative to statistical regression that is used in cases when the model is indefinite or the relationships between model parameters are vague. The possibilistic regression can also be used when the data are hierarchically structured. In biological applications, the hierarchical data structure is inherently present in biodiversity studies implied by species composition and the respective phylogeny. Comparative phylogenetic analyses then correct for the data non-independence, but assume that the observed traits have a significant heritable component. With non-significant phylogenetic signal, the model parameters are more likely to exhibit vague relationships. Such cases justify choice of a possibilistic model, but any possibilistic models are difficult to access for biologists. The possibilistic methods are not implemented in commonly used statistical software. To provide computational tools for biological applications, we implemented five fuzzy linear regression methods in an R package *fuzzyreg*.

The *fuzzyreg* package provides functions for five fuzzy linear regression methods that differ in the expectations of the input data, outlier handling and the applied possibilistic model. The "lee" method utilizes crisp numbers for the explanatory and response variables and predicts values in form of non-symmetric triangular fuzzy numbers. Methods "diamond", "hung", "nasrabadi" and "tanaka" require a fuzzy response variable, with the "nasrabadi" method accepting also fuzzy explanatory variable values. The package contains functions for algebraic operations with triangular fuzzy numbers and a wrapper function for fitting the fuzzy linear models *fuzzylm()* that outputs an object of the newly defined class *fuzzylm*. Generic functions associated with class *fuzzylm* include *print*, *summary*, *plot* and *predict*. The data in the package contain examples from the original descriptions of the methods and the tests show that our implementation in R outputs results identical to those obtained by the methods' authors.

L1-02

Using Simulations for Informed Design of Experiments

Modrák M.[1]

[1] *Laboratory of Bioinformatics, Institute of Microbiology of the CAS*

In life sciences, it is not uncommon to design experiments without explicitly considering the ensuing computational or statistical processing. This may however lead to experiments that are doomed to failure or, even worse, to experiments that provide misleading results. A notable observation is that if an experiment has low statistical power, many classical statistical tests can claim a "significant" result when the measured effect in fact has the opposite sign than the true effect (Type S error) or is of vastly different magnitude (Type M error) [1]. Calculating statistical power during design of the experiment lets the experimenter notice possible problems early. Nevertheless for non-trivial analytical tools – which are common in bioinformatics – power calculations become challenging to perform and may require expert statistical knowledge.

An accessible but powerful method to understand what can be expected from the experiment is to simulate multiple datasets across the range of plausible effect sizes, run the planned procedure and examine the distribution of results. This process allows the experimenter to estimate statistical power and foresee possible problems without mathematical analysis of the statistical/computational procedure. While most useful for experiment design, simulation studies can also help us better understand the tools we use, interpret experiments after they have been performed and warn us of possible Type S and Type M errors in published literature.

In this talk I will discuss some of the underlying statistical concepts and show two example simulation studies using the classical t-test and the popular differential expression tool DESeq2. Code for the simulations can be found at https://github.com/cas-bioinf/statistical-simulations.

References

[1] Gelman, A. and Carlin, J., 2014. Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9(6), pp.641-651.

L1-03

MS-DIAL and MS-FINDER: Let's Make Metabolomics Data Processing Great Again!

Čajka T.[1,2]

[1] *Institute of Physiology CAS, Department of Metabolomics, Videnska 1083, Prague, 14220, Czech Republic*
[2] *West Coast Metabolomics Center, University of California, Davis, 451 Health Sciences Drive, Davis, California, 95616, United States*

Over the last decade, mass spectrometry-based metabolomics and lipidomics have become established as the key platforms for comprehensive profiling of low-molecular-weight compounds in complex biological systems. However, the processing of MS raw data, such as feature detection, peak alignment, exclusion of false-positive peaks, and automated annotation of peaks based on MS/MS library search is still challenging. In addition, it is estimated that in untargeted metabolomics only 20% of detected molecular features can be identified based on mass spectra library matches. Here, two novel software programs, MS-DIAL and MS-FINDER, will be presented for improved data processing. MS-DIAL is a program for untargeted metabolomics that supports multiple instruments and MS vendors. It features spectral deconvolution, streamlined criteria for peak identification, support of all data processing steps from raw data import to statistical analysis, and user-friendly graphic user interface [1]. MS-FINDER is a program for compound annotation that supports EI-MS (GC-MS) and MS/MS spectral mining. MS-FINDER aims to provide solution for formula predictions, fragment annotations, and structure elucidations by means of unknown spectra. In addition, the program can annotate unknowns by using a combination of 14 metabolome databases such as HMDB, PubChem, and UNPD [2].

References

[1] H. Tsugawa, T. Cajka, T. Kind, Y. Ma, B. Higgins, K. Ikeda, M. Kanazawa, J. VanderGheynst, O. Fiehn, M. Arita: MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods* 12 (2015) 523–526.

[2] H. Tsugawa, T. Kind, R. Nakabayashi, D. Yukihira, W. Tanaka, T. Cajka, K. Saito, O. Fiehn, M. Arita: Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical Chemistry* 88(16) (2016) 7946–7958.

L1-04

Metabolite Mapper (MM2) - Complex LC/GC HRMS tool for metabolomics data processing

Fesl J.[1,2], Doležalová M.[1], Moos M.[1], Šimek P.[1]

[1] *Laboratory of Analytical Chemistry and Metabolomics, Biology Centre – CAS, Branišovská 31, České Budějovice*
[2] *Institute of Applied Informatics, Faculty of Science, University of South Bohema, Branišovská 31, České Budějovice*

The processing of the metabolomics oriented data is a very complex problem, which requires dedicated tools for its solution. MM2 is the one of such tools. The tool provides all necessary functions like chromatogram deconvolution, peak detection, peak alignment, deisotoping, dynamic time warping, blank subtraction and many more.

The greatest advantage of this solution is the ability of raw vendor data formats reading, which is necessary for the fast and efficient processing. The tool is proposed as the modular platform and is able to work in the parallel and cluster mode.

For the best results evaluation and reproduction, MM2 contains a special quality check control module that allows dynamicly identify only such metabolites, which have the stable changing profiles. Such way allows to rapidly reduce the number of metabolites, which could become potential markers.

MM2 is now a mature platform which has been developed more then 10 years within the direct cooperation of the chemistry and informatics experts.

L1-05

PROFREP and DANTE: Repetitive elements annotation tools for genome assemblies

Hoštáková N.[1], Novák P.[1], Neumann P.[1], Macas J.[1]

[1] *Biology Centre CAS*

Despite its high abundance, repetitive DNA still represents one of the least characterized parts of eukaryotic genomes. We developed new tools which utilize output from RepeatExplorer pipeline to perform localization and classification of repetitive elements on DNA sequences in a scale of whole assembled genomes. The main PROFREP (PROFiles of REPeats) tool uses sequence similarities to repeat libraries identified and characterized by RepeatExplorer to create repetitive profiles of individual repeat classes. Besides determining regions of repeats it also provides their quantitative representation. DANTE (Domain-based ANnotation of Transposable Elements) tool searches for conserved protein domains encoded by various types of mobile elements. Since this detection is more sensitive than search for nucleotide sequence similarities, it is used to complement results obtained by PROFREP. DANTE is run as a part of the PROFREP pipeline, but it can also be executed separately in order to extract sequences of identified protein domains for downstream analysis.

This work was supported by ELIXIR CZ research infrastructure project (MEYS Grant No: LM2015047) including access to computing and storage facilities.

L1-06

Bioinformatics platform for routine diagnostics of chronic lymphocytic leukemia patients

Reigl T.[1], Stranska K.[1,2,3], Pal K.[1,3], Bystry V.[1], Krejci A.[1,4], Pospisilova S.[1,2], Darzentas N.[1,5], Plevova K.[1,2,3]

[1] *CEITEC, Masaryk University Brno, Czechia*
[2] *Department of Internal Medicine – Hematology and Oncology, University Hospital Brno*
[3] *Medical Faculty, Masaryk University, Czechia*
[4] *RECAMO, Masaryk Memorial Cancer Institute, Brno, Czechia*
[5] *University Hospital Schleswig-Holstein, Kiel, Germany*

Introduction: Chronic lymphocytic leukemia (CLL) is an overall heterogeneous disease with varying clinical outcome. However, subsets of patients with highly similar, stereotyped B cell receptors (BcR) share biological and clinical characteristics, with far-reaching implications for personalised care. As BcR sequence analysis is integral in routine CLL diagnostics, there is a need for user-friendly but robust bioinformatic methods for data management, analysis, interpretation, and reporting.

Methods: Our solution mainly uses different components of our 'Antigen Receptors Research Tool' / ARResT bioinformatics platform (bat.infspire.org/arrest) integrated into ARResT/Interrogate: Sanger sequence analysis with genomePD/GLASS, sequence annotation with IMGT/V-QUEST, analysis of patient subsets with the tool ARResT/AssignSubsets and the database ARResT/Subsets. It also allows the user to access stored sequence and biomedical data to put a new case in context, and finally to create informative and consistent clinical reports. Importantly, the platform is also capable of handling high-throughput data, making it a versatile solution.

Results: We have built a prototype, codenamed 'CLLpedia', for a cohort of >2000 CLL cases tested in the University Hospital Brno. CLLpedia combines different, originally stand-alone components for BcR sequence analysis into one powerful and easy-to-use platform for every-day clinical application.

Acknowledgements: Supported by MZCR-AZV 16-34272A, MZCR-RVO 65269705, and MUNI/A/0968/2017.

L1-07

An Automated Design of Thermostable Multiple-Point Mutants: Presenting FireProt

Musil M.[1,2,3], Stourac J.[1,2], Bendl J.[1,2,3], Brezovsky J.[1,2,], Prokop Z.[1,2], Zendulka J.[3], Martinek T.[3], Bednar D.[1,2], Damborsky J.[1,2]

[1] Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic
[2] International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno, Czech Republic
[3] Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

Proteins are increasingly used in numerous biotechnological application. Unfortunately, naturally occurring proteins cannot usually withstand the harsh industrial environment, since they are evolved to function in mild conditions. Therefore, there is a continuous interest in increasing protein stability to enhance their applicability in the wide range of industrial and medical tasks. The plethora of *in silico* tools for the prediction of the effect of amino acid mutations on protein stability have been developed recently. However, only single-point mutations with a small impact on protein stability are typically predicted with the existing tools and have to be followed by laborious protein expression, purification, and characterization. A much higher degree of stabilization can be achieved by the construction of the multiple-point mutants.

Here, we present FireProt[1,2]: a web server for automated design of thermostable multiple-point mutants that utilizes both structural and evolutionary information. FireProt integrates sixteen bioinformatics tools, including force field calculations via FoldX and Rosetta. Two well-established protein engineering strategies are used to design highly reliable thermostable proteins with the usage of the energy- and evolution-based approaches. Furthermore, multiple-point mutants are checked for the potentially antagonistic effects in the designed protein structure. To deal with high time demands of the force field calculations, the original FireProt method was accelerated by the orders of magnitude via the utilization of the smart knowledge-based filters, protocol optimization, and effective parallelization. The server is complemented with an interactive, easy-to-use interface that allows users to directly analyze and further modify designed thermostable proteins. The method was thoroughly experimentally validated and provided for example highly valued hyperstabilized fibroblast growth factor[3]. The server is freely available at https://loschmidt.chemi.muni.cz/fireprot/.

L1-08

3DPatch: fast sequence and structure conservation annotation in a web browser

Jakubec D.[1,2], Vondrášek J.[1], Finn R. D.[3]

[1] Department of Bioinformatics, Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, 166 10 Prague 6, Czech Republic
[2] Department of Physical and Macromolecular Chemistry, Faculty of Science, Charles University, 128 43 Prague 2, Czech Republic
[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Amino acid residues manifesting high levels of conservation are often indicative of functionally significant regions of protein structures. Residues critical for protein folding, hydrophobic core stabilization, intermolecular recognition, or enzymatic activity often manifest lower mutation rates compared to the rest of the protein. Quantitative assessment of residue conservation typically involves querying a sequence against a database, finding similar sequences, aligning them to bring equivalent positions into register, and applying an information theory-based measure to individual columns in the multiple sequence alignment. Understanding how the sequence conservation profile relates in 3D requires its projection onto a protein structure, which can be a time-consuming process.

We developed 3DPatch, a client-side web application that simplifies the task of calculating protein sequence information content, 3D structure identification, and conservation level-based mark-up. 3DPatch utilizes the power of profile hidden Markov models and speed of HMMER3.1 to provide accurate results in a matter of seconds. It was developed with easy integration into other peoples' websites in mind and supports most modern web browsers. 3DPatch is freely available at http://www.skylign.org/3DPatch/.

L1-09

HotSpot Wizard 3.0: Sequence-based Design of Mutations and Smart Libraries

Sumbalova L.[1,2], Stourac J.[1], Martinek T.[2], Bednar D.[1,3], Damborsky J.[1,3]

[1] *Loschmidt Laboratories, Department of Experimental Biology, Masaryk University, Brno, Czech Republic;*
[2] *IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic*
[3] *International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno, Czech Republic*

HotSpot Wizard is an interactive web server for prediction of amino acid residues suitable for mutagenesis and construction of libraries of mutants using in focused directed evolution. A user specifies a protein of interest and HotSpot Wizard will list the most suitable sites for mutagenesis, so-called hot-spots. There are four different strategies used in the HotSpot Wizard: (i) functional hot-spots represented by highly mutable residues located in the active site pocket or in access tunnels, (ii) stability hot-spots represented by flexible residues, (iii) stability hot-spots represented by consensual amino acids, and (iv) correlated hot-spots represented by pairs of coevolving residues that modulate enzyme activity and selectivity. Altogether, 7 databases and 25 computational tools are used for the analysis and predictions.

Recently, we have developed a new version of HotSpot Wizard 3.0 [1], with several new features which makes it more accessible to broader community. The main new feature is the possibility of entering protein sequence as an input. In previous versions, protein structure was the only possible input, which limited usage of HotSpot Wizard only to structure with solved 3D structure. Users could therefore use only a limited number of proteins as an input or create a model of their structure externally. Newly, search for existing structures in PDB [2] or existing models in Protein Model Portal [3] is performed after entering a protein sequence. If the structure is unknown, HotSpot Wizard conducts its modelling using Modeller [4] or I-Tasser [5]. Structure prediction is very difficult and complex problem and results of modelling are never perfect. It is essential to know how good is the model, before using it for identification of hot-spots or creating the libraries. Therefore, quality assessment of models is part of HotSpot Wizard 3.0, providing different quality metrics using: PROCHECK [6], WHATCHECK [7] and MolProbity [8].

Another new feature is a stability prediction module. Mutagenesis of the protein can significantly influence the protein's stability. Stability predictions upon introduction of mutations can further decrease a number of variants needed for experimental testing.

Users can design single-point or multiple-point mutations and HotSpot Wizard provides a change in free energy between wild-type and the mutant protein using FoldX [9] and Rosetta [10] are used. The web tool is available at https://loschmidt.chemi.muni.cz/hotspotwizard.

References

[1] Sumbalova, L., Stourac, J., Martinek, T., Bednar, D., & Damborsky, J. (2018), HotSpot Wizard 3.0: Web server for automated design of mutations and smart libraries based on sequence input information, *Nucleic Acids Research* (in review).

[2] H.M. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, 28, 235-242.

[3] Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., & Schwede, T. (2013). The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, 2013.

[4] Webb, B., & Sali, A. (2014). Protein structure modeling with MODELLER. *Methods in Molecular Biology,* 1137, 1-15.

[5] Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature Methods*, 12, 7-8.

[6] Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, 26, 283-291.

[7] Hooft, R. W., Vriend, G., Sander, C., & Abola, E. E. (1996). Errors in protein structures. *Nature*, 381, 272-272.

[8] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J. & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D*, 66, 12-21.

[9] Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., & Serrano, L. (2005). The FoldX web server: an online force field. *Nucleic Acids Research*, 33, W382-W388.

[10] Kellogg, E. H., Leaver-Fay, A., & Baker, D. (2011). Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*, 79, 830-838.

SESSION 2

Structures

L2-01

Mol*: Towards a common library for web molecular graphics and analysis tools

Rose A. S.[1], Sehnal D.[2,3,4], Koča J.[2,3], Velankar S.[4], Burley S. K.[1,5]

[1] RCSB Protein Data Bank, San Diego Supercomputer Center, University of California, San Diego, La Jolla, CA 92093, USA
[2] CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
[3] National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic.
[4] Protein Data Bank in Europe (PDBe), European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK.
[5] RCSB Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, USA.

Advances in experimental techniques are providing access to structures of ever more complex and larger macromolecular systems. Web-browser based visualization and analysis of macromolecular structures and associated data represents a crucial step in gaining knowledge from these data. To help streamline this process, we present a project called Mol* (/'mol-star/), a common library and a set of tools for working with macromolecular data sets. The project being developed in collaboration with PDBe, RCSB PDB and CEITEC, and builds on previous experience of the authors (LiteMol Suite and NGL Viewer) and includes modules for data storage, in-memory representation, query languages, UI state management, visualization, and tools for efficient data access. The goal of this talk is to provide an overview of the project, its state and explain some of its building blocks in more detail (e.g., the BinaryCIF format, the molecular query language, or the in-memory representation of molecular assemblies).

L2-02

Bringing together structure related functional annotations

Pravda L.[1], Varadi M.[1], Gutmanas A.[1], Velankar S.[1]

[1] *Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Welcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK*

In order to understand macromolecular structures archived in the Protein Data Bank (PDB), it is essential to take into consideration the biological context of these molecules. There are specialised resources that each provide one or more aspects of the biological context, but it takes significant effort to collect and compare all the information that may be relevant to a specific structure.

PDBe-KB (Protein Data Bank in Europe - Knowledge Base) is a new community driven resource under development by PDBe, which will provide functional annotations for structural data that can be used by the scientific community to answer biological questions. PDBe-KB is a collaborative effort between PDBe and a diverse group of biological resource, structural bioinformatics research teams. This new resource will consolidate older services, such as SIFTS, which focuses on providing seamless mappings between PDB entries and other databases, and data from multiple data enrichment project (e.g. the FunPDBe project), which aims to collect and distribute highly enriched, valuable annotations that create a comprehensive biological context for structural models, effectively bringing structure to biology.

L2-03

Molecular Transport in the view of Structural Bioinformatics & Chemoinformatics

Berka K.[1], Pravda L.[2,3], Sehnal D.[2,3], Navrátilová V.[1], Bazgier V.[1], Juračka J.[1], Toušek D.[1,2], Svobodová Vařeková R.[2], Koča J.[2], Otyepka M.[1]

[1] *Dpt Physical Chemistry, RCPTM, Palacký University Olomouc, CZ*
[2] *NCBR, CEITEC, Masaryk University Brno, CZ*
[3] *PDBe, EMBL-EBI, Hinxton, UK*

Transport of molecules within and between cells is vital for their function. The detailed understanding of molecular transport can help to identify and overcome possible hurdles in drug design and describe biological processes [1]. The transport in the inhomogeneous media (membranes, proteins) is complicated by the interactions with surrounding atoms within different layers of membrane or with amino acid residues along the protein channel. These interactions therefore reflect the structure of the membrane or protein as well as the transporting molecule itself. We present three databases 1) an alpha version of the MolMeDB database [2] showing the interactions of chemicals with membranes including penetration, 2) a database with membrane structures of metabolic cytochromes P450 enzymes that may change transport properties of compounds [3], and 3) the ChannelsDB database [4] that gathers channel data for individual entries in PDB. Finally, we have updated our channel detection application MOLEonline [5] with the ability to detect transmembrane pores and with additional interesting features. Combination of these approaches can enlighten the transport at the molecular level and allow its use in many applications.

References

[1] Di Meo F, et al: In Silico Pharmacology: … *Pharm. Res.*, 111, 471, 2016.
[2] Juračka J, Bazgier V, Berka K: MolMeDB database. http://molmedb.upol.cz
[3] Šrejber M, et al.: Membrane-attached mammalian cytochromes P450: … *J Inorg Biochem*, 183, 117, 2018. http://cyp.upol.cz
[4] Pravda L, et al.: ChannelsDB: … *Nucleic Acids Res*, 46(D1), D399, 2018. http://ncbr.muni.cz/ChannelsDB
[5] Pravda L, et al.: MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Res*, gky309, 2018. https://mole.upol.cz

L2-04

Family-wide annotation and schematic 2D visualization of secondary structure elements

Svobodová Vařeková R.[1,2], Hutařová Vařeková I.[1,2,3], Midlik A.[1,2], Hutař J.[1,2], Moturu T. R.[1], Navrátilová V.[4], Koča J.[1,2], Berka K.[4]

[1] CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
[2] National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
[3] Faculty of Informatics, Botanická 68a, 602 00 Brno-Bohunice, Czech Republic
[4] Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University, 17. listopadu 1192/12, 771 46 Olomouc, Czech Republic

Large amount of structural data is available about individual protein families originating from different organisms, binding various ligands and containing diverse mutations. Secondary structure elements (SSEs) are often characteristic for a protein family and participate in the formation of protein fold. Composition and organization of SSEs can help to identify similarities and differences between proteins within protein families.

To utilize the SSEs for research on protein families, we should be able to easily and automatically link and intercompare corresponding SSEs within one protein family. Specifically, the corresponding SSEs should have the same name (annotation) and we need a transparent and coherent schema of their localization in the protein structures, retaining structural information (i.e., minimize a deviation between SSEs distances in 2D schema and in 3D structure). We developed a tool set [1] fulfilling these requirements. It was successfully tested on all CATH families and its utilization in Protein Data Bank and CATH is planned.

References

[1] Svobodová Vařeková, R., Midlik, A., Hutařová Vařeková, I., Hutař, J., Navrátilová, V., Koča, J., Berka, K. (2018). Secondary Structure Elements-Annotations and Schematic 2D Visualizations Stable for Individual Protein Families. *Biophys J*, 114(3), 46a-47a.

SESSION 3

Cheminformatics

L3-01

African Medicinal Plants: Natural Product Database Development, Lead Discovery and Toxicity Assessment

Ntie-Kang F.[1,2]

[1] *Department of Informatics and Chemistry, University of Chemistry and Technology Prague, Technická 5 166 28 Prague 6 - Dejvice, The Czech Republic*
[2] *Department of Chemistry, University of Buea, P.O. Box 63 Buea, South West Region, Cameroon*

Within the last two decades, drug discovery based on natural products was almost considered as old-fashioned, as combinatorial chemistry quickly took the centre of the stage. However, the decades of combinatorial chemistry, coupled with high throughput screening facilities did not deliver the expected outcomes in terms of new chemical entities and drug approvals. There seems to be a renewed interest in natural product-based discovery, as increasingly new tools are being developed in order to accelerate natural product dereplication and lead discovery, assisted by molecular modeling. The work presented in this thesis focuses on new natural product database tools and datasets for the discovery of lead compounds from African floral matter. Prior to the investigations, data regarding compounds which had been identified from the aforementioned sources were scattered in literature sources, some of which were inaccessible to the wider community of scientists. The resulting investigation has led to a collection of data on the constituent metabolites, their biological activities, as well as the uses of the source organisms in traditional medicine, which have been made available via the web. Moreover, the investigations have led to the identification (assisted and non assisted by molecular modeling) of lead compounds with anti-HIV, anti-*Onchocerca*, antiplasmodial, protease inhibitory and sirtuin inhibitory properties, beginning from plants with popular uses in African traditional medicine. In parallel, the natural product-inspired discovery of antiplasmodial and sigma-binding new chemical entities has been described. Another aspect of the investigations involved the prediction of the drug metabolism and pharmacokinetics of the secondary metabolites, as well as an overall toxicity assessment of the compounds and databases developed, including the development of a new knowledge base for toxicity prediction. The results presented in this thesis constitute the first outcome of the computer-based investigation of the potential of African medicinal plants for drug discovery.

Keywords: African flora, compound libraries, docking, *in silico*, lead compounds, natural product database, pharmacophore, QSAR, virtual screening.

L3-02

Probes & Drugs portal: an interactive, open data resource for chemical biology

Skuta C.[1], Popr M.[1], Muller T.[1], Jindrich J.[1,2], Kahle M.[1], Sedlak D.[1], Svozil D.[1,3], Bartunek P.[1]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[2] Department of Organic Chemistry, Faculty of Science, Charles University, Prague, Czech Republic
[3] Laboratory of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic

Chemical probes became indispensable tools in modern biology. These small molecules are used to study the gene function, validate molecular targets and dissect complex processes within the cells and organisms. Although, a probe ought to be a well-described, potent, selective tool, it has been shown that not all compounds developed/presented as probes possess these qualities. Unfortunately, not only probes, but many other bioactive tools with long disproved biological properties often still bear the old incorrect descriptions and still are the members of pre-picked commercial screening libraries. On top of that, these libraries, also together with many non-commercial ones, are often very hard to acquire in a computer-readable form which would enable to analyze them using standard cheminformatics tools. In order to make these available, to enable literally anyone to work with them and ask from simple to very complex, multi-conditional questions in the context of compounds commonly used in chemical biology, we present the Probes & Drugs (P&D) portal (https://www.probes-drugs.org). P&D portal is a tool for the exploration of bioactive compound space from various points of view through an intuitive and yet very powerful filtering system. This system, enhanced by the Boolean logic, and in a combination with integrated visualizations, ontologies and chemical intelligence makes it a unique discovery platform with practically an unlimited number of query possibilities. P&D portal is an up-to-date web resource that joins the worlds of commercial and public bioactive compound sets which, via its unparalleled filtering system, enables to resolve diverse logical queries unlike any similar resource.

L3-03

Sachem: A chemical cartridge for high-performance substructure search

Galgonek J.[1], Kratochvil M.[1,2], Vondrášek J.[1]

[1] *Institute of Organic Chemistry and Biochemistry of the CAS, Flemingovo nám. 2, 166 10 Praha 6*
[2] *Faculty of Mathematics and Physics, Charles University, Malostranské nám. 25, 118 00 Praha 1*

Structure search is a valuable approach for exploring small-molecule databases. Fingerprint-based screening methods are usually employed to enhance the search performance by reducing the number of calls to the verification procedure. In substructure searches, fingerprints are designed to capture important structural aspects of the molecule to aid the decision about whether the molecule contains a given substructure. Most currently available cartridges provide acceptable search performance for processing user queries, but do not scale satisfactorily with dataset size.

We present Sachem, a new open-source chemical cartridge that implements two substructure search methods. The first is a performance-oriented reimplementation of substructure indexing based on the OrChem fingerprint, and the second is a novel method that employs newly designed fingerprints stored in inverted indices. We assessed the performance of both methods on small, medium, and large datasets containing 1 million, 10 million, and 94 million compounds, respectively. Comparison of Sachem with other freely available cartridges revealed improvements in overall performance, scaling potential, and screen-out efficiency.

The Sachem cartridge allows efficient substructure searches in databases of all sizes. The sublinear performance scaling of the second method and the ability to efficiently query large amounts of pre-extracted information may together open the door to new applications for substructure searches.

L3-04

Advances in interpretation of QSAR models

Polishchuk P.[1]

[1] *Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University and University Hospital in Olomouc, Hněvotinska, 1333/5, Olomouc, Czech Republic*

Starting from very beginning methods establishing structure-activity relationship were rather simple and obtained models had straightforward and clear interpretation. This was frequently used to explain observed relationships that can be used for design of compounds with desired properties. The first QSAR models used linear regression or decision tree algorithm were interpretable but had low predictivity and robustness. Therefore, more sophisticated machine learning methods like random forest, support vector machine and neural networks quickly became popular for QSAR modeling because they result in highly predictive models. Those machine learning models were hardly interpretable and frequently were considered as “black boxes”. There are several approaches which can solve this issue to calculated descriptor contributions from any kind of models including those ones. These are sensitivity analysis, partial derivatives and importance measurement. However, the remaining limitation is the interpretability of descriptors used for modeling. This limitation was overcome by recently developed approaches that make all models interpretable irrespective to machine learning methods and descriptors used. They all utilize the idea of matching molecular pairs. The contribution of a fragment is calculated as a difference between the predicted activity for the initial compound bearing the fragment and the virtual compound with the removed fragment of interest. Calculated fragment contributions are context dependent. Therefore, it is important to take this into consideration for model interpretation and recent developments solve this issue. Calculated contributions of fragments can be used for revealing of structure-activity relationship trends or to determine directions of structure optimization. However, there is still a need to develop user-friendly programs which can be used by non-experts.

SESSION 4

Sequences

L4-01

Isometric gene tree reconciliation revisited

Brejová B.[1], Chládek R.[1], Gafurov A.[1], Pardubská D.[1], Sabo M.[1], Vinař T.[2]

[1] *Department of Computer Science, Faculty of Mathematics, Physics, and Informatics, Comenius University, Bratislava, Slovakia*
[2] *Department of Applied Informatics, Faculty of Mathematics, Physics, and Informatics, Comenius University, Bratislava, Slovakia*

Isometric gene tree reconciliation is a gene tree / species tree reconciliation problem where both the gene tree and the species tree include branch lengths, and these branch lengths must be respected by the reconciliation. The problem was introduced by Ma et al. 2008 in the context of reconstructing evolutionary histories of genomes in infinite sites model. We have shown that the original algorithm by Ma et al. is incorrect, and we propose a modified algorithm that addresses the problems that we discovered. Our algorithm is also more efficient. We have further studied several new variants of the problem, including reconciliation of two unrooted trees and allowing uncertainty in the input branch lengths.

Some of the results have apeared in B. Brejová, A. Gafurov, D. Pardubská, M. Sabo, T. Vinař. Isometric gene tree reconciliation revisited. Algorithms for Molecular Biology, 12:17. 2017.

L4-02

Analyzing Raw Signal From MinION Sequencers

Rabatin R.[1], Brejová B.[1], Nosek J.[2], Vinař T.[1]

[1] *Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava*
[2] *Faculty of Natural Sciences, Comenius University in Bratislava*

One of the great advantages of MinION sequencing is its ability to directly observe base modifications, such as methylation, in the DNA sample. The base modifications cause systematic shifts in the current measured when DNA strand is passing through the pore. These shifts can be later identified in software. Current tools, such as Tombo, typically require sequencing of matched control samples, where DNA modifications were removed, e.g. by in vitro synthesis using PCR. The signal obtained by sequencing the control sample is compared to the signal obtained from the sample with modifications using a statistical test. Another option is to use models that characterize signal progression for known modifications; such models are estimated from data obtained through artificial methylation. Our goal is to develop an usupervised modification detection framework. In our work, we first use deep machine learning techniques to train characteristics of a typical signal from samples where DNA modifications were removed. Next, we scan newly sequenced samples for sections that do not conform to the learned model by using anomaly detection techniques. The advantage of this approach is that it is sufficient to train the model once for a particular chemistry, and consequently we do not require control samples matched to the analyzed sequence. Currently, we are working on identifying known methylation patterns, but the framework can be naturally extended towards truly unsupervised modification detection, with potential of detection of previously uncharacterized modifications.

L4-03

G-quadruplex forming sequences in nucleic acids and their detection in-silico

Lexa M.[1], Hon J.[2], Martínek T.[2]

[1] *Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno*
[2] *Faculty of Information Technology, Brno University of Technology, Božetěchova 2, 612 66 Brno*

G-quadruplexes are non-canonical DNA or RNA structures with four strands or strand segments interacting via guanine hydrogen bonds (Hoogsteen bonds). Biologically interesting quadruplexes are typically formed by a single nucleic acid strand folded onto itself with guanines present in four distinct (in complicated cases only partially distinct) runs of Gs. Sometimes G-quadruplexes formed by two strands are also considered. While these structures have long been known and telomere quadruplexes have been studied for decades, they have come under increased scrutiny only recently with interest in their genome-wide presence. This was a logical result of numerous fresh genome sequencing and annotation efforts as well as discoveries that G-quadruplexes regulate DNA replication, transcription and some additional biological processes in cells. Gene promoters, transposable elements and virus genomes were found to be enriched in G-quadruplex forming sequences.

The increased interest also requires a better ability to search genomic sequences for sequences with high potential for G-quadruplex formation. The existing approaches range from initial use of simple patterns to more recent methods based on machine learning. We review past and present tools and approaches to G4 detection in-silico, including our own original work in this area that resulted in a tool called pqsfinder.

L4-04

Shedding light on the “index hopping” problem in Illumina sequencing technology for amplicon data

Morais D. K.[1], Baldrian P.[1], Větrovský T.[1]

[1] *Institute of Microbiology of the Czech Academy of Sciences, Videnska 1083, 14220 Prague 4, Czech Republic*

High-throughput sequencing technologies revolutionized the way biologists deal with data and introduced the biological field to the big-data world. These tools allowed microbial ecologists to assess microbial communities in a level of depth never seen before. However, at the same time, we got analytical and technological problems we have never faced before. For example, the “index hopping” (also called as “tag switching”) events observed in Illumina sequencing technology when using a multiplexed approach. The phenomenon was recognised by the fact that sequence barcodes (also called indexes or tags) were found to be exchanged between samples during the sequencing run forming even such forward and reverse barcode combinations that were not used during amplicon preparation. Furthermore, if a forward or reverse barcode is used for multiple samples during PCR, index hopping can generate incorrect reads with “correct” (existing) barcode combination. Until now, this phenomenon has only been described for 454 and Illumina HiSeq platforms and there are no tools focused on this specific matter. This research aims to supply a tool to identify the switched barcodes and to quantify the frequency of these erroneous events, shedding light on this type of problem. We developed an R script that allows users to identify their sequences by the barcode pair used for each sample, to detect and quantify all the unexpected barcode pairs in a paired-end Illumina sequencing result. The script was completely written using base R functions to avoid unstable dependencies, with most of the functions in vectorised forms to improve speed and memory consumption. The script can be downloaded from https://github.com/kdanielmorais/Tag_counting_barcodes. We have inspected five Illumina MiSeq libraries sequenced with the V3 kit that contained 20 barcode pairs and we detected an average of 5% index hopping. For this dataset, the hopping frequency seemed to be affected by barcode sequences or numbers of barcodes used. We believe that a bigger dataset might highlight other important factors and help to explain the molecular mechanisms causing index hopping.

L4-05

Assembling diploid genome of Cobitis taenia using Illumina short reads

Mokrejš M.[1], Bartoš O.[2,3], Röslein J.[2,4], Kočí J.[2,4], Janko K.[2,4]

[1] *IT4Innovations, VŠB – Technical University of Ostrava, Ostrava, Czech Republic*
[2] *Institute of Animal Physiology and Genetics, Laboratory of Fish Genetics, The Czech Academy of Sciences, Liběchov, Czech Republic*
[3] *Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic*
[4] *Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czech Republic*

Spine loaches (*Cobitidae*) are common fishes in small rivers and creeks in Eurasia. We determined genome sequence of *Cobitis taenia*, denoted as tt_16D1C3L12 strain. The whole *Cobitis* species-hybrid complex is interesting because its hybridogenetic forms reproduce asexually through gynogenesis and therefore, our research might shed light on the „paradox of sex“ [1].

We obtained 2x250bp long reads from Illumina HiSeq 2500 with median insert sizes xx and xx, respectively. We also obtained Nextera long mate-pair reads for several libraries with media insert sizes xx, xx and xx bp. Based on k-mer size analysis of trimmed and error-corrected reads the diploid genome appears to be around 1.267 Gbp. The homozygous loci were covered almost 40x and heterozygous loci have half-coverage (about 20x). The genomic sequence was assembled using abyss-2.0.2 into 320183 contigs (>=500bp) using solely paired-end read data and further scaffolded using Illumina long-mate pair reads into 195685 scaffolds (>=500bp). We show qualities of Nextera long mate-pair libraries. We further report performance of SPAdes-3.11.1 assembler on the same input dataset. Similarly, we report performance of Pilon and Sealer gap-closing programs. From preliminary analyses we conclude only about 34% of short gaps were at least partially filled in. In the case of such in-filled gaps, mostly only 90% of each gap was filled in. This work was also partially supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center – LM2015070".

References

[1] Bell, G. (1982). The Masterpiece of Nature: The Evolution and Genetics of Sexuality. CUP Archive.

L4-06

Detection of selection pressure in human genome

Ehler E.[1], Pačes J.[1], Moravčík O.[1]

[1] *Department of Genomics and Bioinformatics, Institute of Molecular Genetics of the Academy of Sciences of the Czech Republic*

In our talk, we will present our research of selection detection in human genome. As a pilot study, we have investigated the traces of selection around forensic short tandem repeat (STR) loci. Main objective of the study was to test the selection detection methodology and approaches to be used on larger datasets and projects, namely the human endogenous retroviral (HERV) loci. In conjunction with this objective we developed a new method, the *ExP heatmap*, for visualization of the multidimensional cross-population results. Second objective was to identify the STR loci where the distribution of alleles in population could be affected by selection sweeps. Third objective involved evaluation of inter-population differences in selection pressure.

We gathered publicly available data from The 1000 Genomes Project (phase 3). The sample-set consisted of 2,504 unrelated individuals. For the total number of 24 forensic loci, we gathered the single nucleotide polymorphisms (SNP) haplotypes spanning one million base pairs (Mbp) window around the STR locus. Fifty additional one Mbp regions were used as comparative data (EDAR, LCT, SLC45A2, and 47 randomly selected). Total number of investigated SNPs was 182,601. To assess the selection pressure at these loci, we have computed descriptive statistics (heterozygosity, derived allele frequency, Hardy-Weinberg equilibrium deviation), as well as eleven selection tests: Tajima's D, iHS, nSL, Garud's H1, Garud's H2/H1, LKT, LKT with island model, LKT with hierarchical island model, FST, XP-EHH, and XP-nSL.

With our results, we can confirm different modes of selection acting on human populations in the 1000 Genomes dataset. From the tested group of 24 forensic microsatellites, we detected selection signal at minimal eight loci (D3S1358, D5S818, D10S1248, D18S51, D22S1045, TH01, TPOX, and vWA). No selection signals were uncovered from D6S1043, D8S1179, D13S317, D19S433, D21S11, and Penta D loci. The selection signals proved to be significantly different between various populations and population clusters (super-populations). We will also report on preliminary results of detecting selection on HERVs and integrating our results into HERVd database (https://herv.img.cas.cz).

L4-07

Detecting natural selection signal in bat DNA sequences after exposure to white-nose syndrome

Harazim M.[1,2], Martínková N.[1,3]

[1] *Institute of Vertebrate Biology, Czech Academy of Sciences, Brno, Czech Republic*
[2] *Department of Botany and Zoology, Faculty of Science, Masaryk University, Brno, Czech Republic*
[3] *Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic*

Two populations with spatio-temporal differences in exposure to a pathogen vary in the response to the infection. With increasing number of lethal cases, the infection imposes a strong selective pressure, which lead to adaptations detectable at the genomic level. Positive selection is defined as $d_N / d_S > 1$, where d_N represents non-synonymous substitution rate and d_S synonymous substitution rate in DNA sequences. Positive selection signal in specific genes with function related to the infection progress provides an insight to the mechanism of disease tolerance or resistance.

We studied white-nose syndrome (WNS), a fungal infection of hibernating bats, that is tolerated by the Palearctic bats, but lethal in multiple Nearctic bat species. Palearctic bats represent populations with historic exposure to the pathogen, and we hypothesized that past selective pressure of the pathogen resulted in genomic changes promoting infection tolerance.

We sequenced selected genes with function in water metabolism and skin structure on the Pacific Biosciences platform. We investigated partial sequences of 23 genes in nine Palearctic and Nearctic hibernating bat species and one non-hibernating species for signal of natural selection in a phylogenetic context. With maximum likelihood analysis, we found that eight genes were under positive selection, and we successfully identified amino acid sites under selection in five encoded proteins. The branch site models revealed positive selection in three genes.

Palearctic bats exhibit signals for positive selection in genes with functions ensuring cell membrane fluidity with changing temperature, tissue regeneration and wound healing, and also modulation of the immune response. We developed a mechanistic model that highlights the importance of skin barrier integrity and healing capacity in the progression of WNS pathophysiology and propose a possibility of downregulation of the immune reaction in response to the *Pseudogymnoascus destructans* infection.

Poster Session

Monday, 11. June

Poster session is sponsored by the
National Infrastructure of Chemical Biology and
Protein Data Bank in Europe.

P-01

Study of GR morphs using conformal classification to define applicability domain

Agea M. I.[1], Svozil D.[1,2]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic
[2] CZ-OPEN-SCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic

Defining the prediction boundaries of an in silico model is as important as determining its statistical quality. Applicability domain (AD) is defined by the training set objects and it refers to the region of space in which new predictions can be made. The aim of this study is to show that our molecular morphing algorithm [1] is able to produce a virtual library enriched by actives against a selected target, as compared to a randomly generated library. Particularly, the glucocorticoid receptor morphing library (GRML), generated from known actives of the GR receptor, was compared to the library generated from randomly selected ZINC compounds (RML). Morphs in each library were predicted to be active or inactive at the GR receptor by random forest classification with AD defined using conformal prediction (CP). We show, that the GRML contains 75-times more ligands predicted (number of predicted actives=44170) to be active than the RML (number of predicted actives=587). Furthemore, we built QSAR regression model to predict potencies of predicted active morphs. Finally, we projected predicted active morphs into ZINC database in order to obtain the list of structurally-related purchasable compounds.

References

[1] Hoksza D, Skoda P, Voršilák M, Svozil D. Molpher: a software framework for systematic chemical space exploration. J Cheminform. 2014 Mar 21;6(1):7.

P-02

Genomic (un)stability in hybridogenetic clonal forms of European loaches (genus Cobitis)

Bartoš O.[1,2], Röslein J.[1,3], Kočí J.[1,3], Mokrejš M.[4], Janko K.[1,3]

[1] *Institute of Animal Physiology and Genetics, Laboratory of Fish Genetics, The Czech Academy of Sciences, Liběchov, Czech Republic*
[2] *Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic*
[3] *Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czech Republic*
[4] *IT4Innovations, VŠB – Technical University of Ostrava, Ostrava, Czech Republic*

Although molecular and cytological mechanisms underlying meiosis are highly conserved they have been disrupted in numerous ways during the evolution leading to emergence of so called asexual lineages [1]. The origin and mainly persistence of asexual organisms represents major challenge for evolutionary biology [2]. Interestingly, quite recently it has been reported that asexual *Daphnia* strains lose their genes (DNA) in a pace of cca. two genes per generation, which has been interpreted as possible/suitable solution to so called "paradox of sex" [3].

Previous cytogenetic study of Majtánová *et al.* [4] reported no detectable chromosomal rearrangements concerning the European loaches asexual hybrid lineages. So, it seems that species-specific chromosomes are passed from generation to generation exactly in the same form as inherited from the parental species in the time of clonal lineage establishment (clonality). However, will this overall impression of genomic stability persist on "fine" genomics scale?

In this study we have evaluated 46 fish samples - parental species as well as their hybrids. We applied exome capture pair-end sequencing to discover species specific SNPs and determined their coverage (GATK 3.8, samtools 0.1.19, R 3.3.2 and custom python scripts). Which allowed us to identify all presumably Loss Of Heterozygosity loci. Suggesting that gene conversion or gene loss are the processes that stand behind LOH, we ask what is their relative contribution to the overall pattern. To distinguish between those two processes, we utilize the basics of logic used in rna-seq (DifferentialExpression) experiments.

We found out that the hypothesis of "clonality" roughly holds. Nevertheless, we have identified significant amount of interplay between the parental genomes which is manifested/detected as Loss Of Heterozygosity. Further we demonstrate that amount of such interplay is (evolutionary) time-dependent.

This work was also partially supported by The Ministry of Education, Youth and Sports from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070".

References

[1] Bengtsson, B. O. (2009). Asex and evolution: a very large-scale overview. In Lost sex (pp. 1-19). Springer, Dordrecht.

[2] Bell, G. (1982). The Masterpiece of Nature: The Evolution and Genetics of Sexuality. CUP Archive.

[3] Tucker, A. E., Ackerman, M. S., Eads, B. D., Xu, S., & Lynch, M. (2013). Population-genomic insights into the evolutionary origin and fate of obligately asexual Daphnia pulex. Proceedings of the National Academy of Sciences, 110(39), 15740-15745.

[4] Majtánová, Z., Choleva, L., Symonová, R., Ráb, P., Kotusz, J., Pekárik, L., & Janko, K. (2016). Asexual reproduction does not apparently increase the rate of chromosomal evolution: karyotype stability in diploid and triploid clonal hybrid fish (Cobitis, Cypriniformes, Teleostei). PloS one, 11(1).

P-03

StReC: Tool for Prediction of Selectivity of Steroid Receptors

Bazgier V.[1], Otyepka M.[1], Berka K.[1]

[1] *Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacky University, 17. Listopadu 12, 771 46 Olomouc, Czech Republic*

Steroids as hormones play multiple roles in the regulation of human body via binding to individual steroid receptors. We present new user-friendly web tool for prediction of binding of selected candidate molecules into the matrix of human steroid receptors. The results may predict the selectivity of candidate new designed compounds between individual steroid receptors. The beta application is available free for use at http://strec.upol.cz after the registration.

This work was funded by the LM2015047 project from Ministry of Youth, Education and Sports of the Czech Republic and support by the Student projects IGA_PrF_2018_032 of Palacky University.

P-04

Dante - genotyping of complex and expanded short tandem repetitions

Budiš J.[1,2,3], Kucharík M.[4], Ďuriš F.[2,3], Gazdarica J.[2,5], Zrubcová M.[5], Ficek A.[5], Szemes T.[2,5,6], Brejová B.[1], Radvanszky J.[2,5,7]

[1] *Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*
[2] *Geneton Ltd., Bratislava, Slovakia*
[3] *Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia*
[4] *Medirex a.s., Bratislava, Slovakia*
[5] *Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia*
[6] *Comenius University Science Park, Bratislava, Slovakia*
[7] *Biomedical Research Centre, Slovak Academy of Sciences Bratislava, Slovakia*

Short tandem repeats (STRs) are stretches of repetitive DNA in which short sequences, typically made of 2-6 nucleotides, are repeated several times. Since STRs have many important biological roles and also belong to the most polymorphic parts of the human genome, they became utilized in several molecular-genetic applications. Precise genotyping of STR alleles, therefore, was of high relevance during the last decades. Despite this, massively parallel sequencing (MPS) still lacks the analysis methods to fully utilize the information value of STRs in genome scale assays.

We propose an alignment-free algorithm for genotyping and characterizing STR alleles based on sequence reads originating from STR loci of interest called Dante. The method accounts for natural deviations from the expected sequence, such as variation in the repeat count, sequencing errors, ambiguous bases, and complex loci containing several different motifs.

In addition, we implemented a correction for copy number defects caused by the polymerase induced stutter effect as well as a prediction of STR expansions that, according to the conventional view, cannot be fully captured by inherently short MPS reads.

We tested Dante on simulated data sets and on data sets obtained by targeted sequencing of protein coding parts of thousands of selected clinically relevant genes. In both these data sets, Dante outperformed HipSTR and GATK genotyping tools. Furthermore, Dante was able to predict allele expansions in all tested clinical cases.

The presented work was supported by the “REVOGENE – Research centre for molecular genetics” project (ITMS 26240220067) supported by the Operational Programme Research and Development funded by the ERDF.

P-05

Database tools and other software for the study of transposable elements

Červeňanský M.[1], Lapár R.[2], Jedlička P.[1], Kejnovský E.[2], Lexa M.[1]

[1] *Department of Plant Developmental Genetics, Institute of the Biophysics of the Czech Academy of Sciences, Královopolská 2590/135, 612 00 Brno-Žabovřesky*
[2] *Faculty of Informatics, Masaryk University, Botanicka 68a, 60200 Brno*

Transposable elements are autonomously replicated repetitive sequences present in high numbers in eukaryotic genomes. In evolution of individual species and their genomes these sequences go through repeated rounds of multiplication and reduction, often producing and leaving behind only fragments of full-length sequences. The sequences are also mixed by nesting and recombination as well as mutated with time by normal biological processes. With the number of annotated genomes rapidly increasing, it is becoming very difficult to keep track of all the existing variants and their mutual positions and relationships in genomes. We are developing tools to deal with this kind of data. Specifically, we are building a CHADO (www.gmod.org) based environment for storage and visualization of transposon annotations and developing a software tools for detection of ancient transposon copies fragmented by insertion of younger copies (nesting). Recent progress of our work on these tools will be presented.

P-06

The speech of structural patterns

Čmelo I.[1], Svozil D.[1]

[1] *Laboratory of Informatics and Chemistry, UCT Prague, Technická 5, CZ-166 28 Prague 6, Czech Republic*

Many cheminformatic methods extensively utilize structural fingerprints, and often treat their individual bits (and thus whatever structural features the bits represent) as fully independent variables. However, some structural features are interdependent, either explicitly due to structural overlap between the patterns, or implicitly due to positive or negative interactions between structurally unrelated features within a given set of structures. Therefore, a quantitative feature interdependence analysis could yield information useful for interpreting the results of fingerprint-based methods and for their further improvement.

This poster presents a method to quantify feature interrelations for arbitrary sets of compounds, and demonstrates its use by exploratory analysis of four well-known chemical databases: DrugBank, ChEMBL, PubChem and ZINC. The method is based on pointwise mutual information, a concept from general information theory that is often used in the field of computational linguistics to quantify word interrelations.

P-07

Approximate string matching approaches for genomic data

Cvacho O.[1], Hrbek L.[1], Holub J.[1]

[1] *Faculty of Information Technology, Czech Technical University in Prague*

Approximate string matching is essential operation in many bioinformatics applications. For example, mapping reads from high-throughput sequencing onto reference genome, or in the microarray probe design process. Approximate string matching is a task to find all occurrences of given pattern P in a text T with some maximum number k of mismatches allowed. The number of mismatches is measured using distance metric that expresses the minimum number of operations required to transform one sequence into another (or its substring). We are interested in Hamming and Levenshtein distance. Hamming distance uses only one operation substitution and it is defined for two sequences of the same length. Levenshtein distance is defined as the minimum number of operations substitution, insert and delete.

The key ideas for our proposed algorithm are to use compressed self-index and low memory usage of the index and all additional data structures. Multiple approaches exist for solving approximate string matching. Dynamic programming (Needleman-Wunsch, SmithWaterman), automata-based algorithms, and filtering algorithms.

We focus on solving an approximate string matching problem using filtering technique. The first method is based on the pigeonhole principle. The idea is to divide the pattern into substrings and search for them using exact search. All found locations are examined for a possible occurrence. The second method is a generation of the pattern neighborhood. Neighborhood generation uses traversal of the de Bruijn graph (dBG) to filter out non-perspective pattern variants. dBG is constructed from all k-mers that are present in the input genome.

We compared these two methods against state of art searching application currently used in bioinformatics. We focused on time and memory differences, advantages and disadvantages of each method for short and long patterns and proposed possible combination. We have developed an approximate string matching algorithm, and its variants, that will detect all approximate matches. Results show that neighborhood generation method is able to reduce the number of pattern variants nearly four times and still maintaining competitive running time. Resulting index has smaller memory requirements than Bowtie. The other method shows promise for better running time for large genomes in comparison to BLAST.

P-08

QSAR affinity fingerprints: Further exploration of chemogenomic space

Dehaen W.[1], Škuta C.[1,2], Svozil D.[1,2]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic
[2] CZ-OPEN-SCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic

Chemogenomic space describes the relationship between chemical and genomic space and one logical entrance point to it is the ligand-target space, i.e. the activity of all possible ligands to all possible targets. Most of these activities are empirically not known, leading to a very sparse experimental ligand-target matrix. Using the ChEMBL database of measured target-ligand activities and QSAR we attempt to construct models for each target to fill in this chemogenomic space, and with this modelled ligand-target matrix as a starting point, we construct a series of ligand fingerprints based on QSAR-predicted activity at various targets.

Various aspects of these QSAR affinity fingerprints (QAFFP) will be discussed in this poster, including the predictive value at targets outside of the fingerprint, the effect of target selection to build the fingerprint, the biological relevance of these fingerprints, the effect of different machine learning techniques on the predictive strength of the fingerprints, comparison with structure based fingerprints and combination of QAFFP with structural fingerprints to build a hybrid fingerprint.

P-09

Effect of temperature on DNA structure

Dohnalová H.[1], Daršata T.[1], Lankaš F.[1]

[1] *Department of Informatics and Chemistry, UCT Prague*

DNA is a carrier of all cellular genetic information. Sequence-dependent conformation of the DNA affects its interaction with proteins, gene expression and chromatin organization. Temperature influences the DNA structure, which is important for understanding biological processes in thermophilic organisms. In our work we study the interplay between conformational substates of the DNA backbone, spatial arrangement of its bases and basepairs, and the global shape of the DNA double helix as a function of temperature. To this end we performed atomic-resolution molecular dynamics simulations of a 33 basepair DNA oligomer with explicit inclusion of water and ions at five different temperatures ranging from 7 °C to 47 °C. From the simulated data, we extract temperature dependence for all global and local coordinates defining the DNA geometry. To understand the underlying microscopic mechanism, we investigate a relationship between the coordinate changes and the backbone torsion substates.

P-10

Efficient algorithm for compound elemental composition prediction in biological matrices

Doležalová M.[1], Fesl J.[1,2], Moos M.[1], Šimek P.[1]

[1] *Biology centre CAS, Branišovská 31, České Budějovice, 37005*
[2] *University of South Bohemia, Faculty of Science, Institute of Applied Infromatics, Branišovská 31a, České Budějovice, 37005*

The identification of unknown compounds in biological matrices is a very frequent task. The accurate masses measured by the LC HRMS devices allows us to advertise probable summary formulas of such compounds. The increasing mass but means the increasing count of advertised formulas.

There have been published some strategies which allows to reduce the number of generated formulas. These strategies are mostly based on heuristics rules - like specific elemental counts, ratios etc.

Our team havedeveloped a new algorithm summarizing the knowledges of elemental counts got from the HMDB and KEGG databases and serves for better identification of unknown compounds.

P-11

Size-inferred analysis of fetal and maternal signals in noninvasive prenatal testing

Budiš J.[1,2,3], Ďuriš F.[2,3], Radvanszky J.[2,5,7], Kucharík M.[4], Szűcs G.[1], Striešková L.[5], Haršanyová M.[5], Minárik G.[4], Sekelská M.[4], Turňa J.[3,6], Szemes T.[2,5,6]

[1] *Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*
[2] *Geneton Ltd., Bratislava, Slovakia*
[3] *Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia*
[4] *Medirex a.s., Bratislava, Slovakia*
[5] *Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia*
[6] *Comenius University Science Park, Bratislava, Slovakia*
[7] *Biomedical Research Centre, Slovak Academy of Sciences Bratislava, Slovakia*

Objectives: Nowadays, noninvasive prenatal testing (NIPT) is an integral component of obstetric practice. The primary aim of prenatal testing is screening for fetal aneuploidies such as trisomy of chromosome 21 (Down syndrome) or monosomy X (Turner syndrome). However, signals from maternal DNA aberrations such as mosaicism, copy number variations, duplications or deletions can be attributed to fetus, thus resulting in false positive or false negative results. Therefore, it is important for NIPT to reliably detect such artefacts.

Methods: We developed a method to distinguish maternal and fetal signals in NIPT results. The method is based on the length of circulating cell free DNA fragments, which is a mixture of both fetal and maternal origin. The method is purely computational in nature and does not require any additional data apart from those obtained through regular NIPT (such as from trisomy detection).

Results: We tested the presented method on real samples collected for NIPT of common fetal aneuploidies. We showed that the method could detect all positive (i.e., trisomic) fetal signals as well as significantly reduce the number of uninformative results.

Conclusions: The presented method demonstrated that it is possible to distinguish between fetal and maternal signals in NIPT results. This method is thus able to reduce the number of false positive, false negative and uninformative results which further improves the quality of the rapidly advancing NIPT.

P-12

Genome-wide identification of meiotic non-crossovers in mice

Gergelits V.[1], Parvanov E.[1], Simecek P.[2], Forejt J.[1]

[1] *BIOCEV, Institute of Molecular Genetic, ASCR, Prague*
[2] *The Jackson Laboratory, Bar Harbor, Maine, USA*

At leptotene/zygotene stage of meiosis I SPO11 induces approximately 250 DNA double-strand breaks (DSBs) per cell in mouse. All DSBs have to be repaired: either by crossover resolution (CO, ~10%) with a mutual exchange of chromatid arms, or non-crossover (NCO, ~90%) resolution with no exchange of chromatid arms. The nonreciprocal recombination begins at DSBs by digestion of single strand 5'end, leaving the 3'overhang strand to invade the homologous chromosome, which is used as a template to extend the 3'strand. This process, also known as gene conversion, is supposed to be essential for starting the homologous chromosomes synapsis. However little is known about localization and length of the sequence copied from the homologous chromosome during these events.

We used our mouse C57BL/6J-Chr #PWD chromosome substitution strains (abbreviated B6.PWD-Chr#) to directly identify and characterize products of the NCO resolution in a chromosome/genome-wide manner. We sequenced whole genomes of six chromosome substitution strains and compared them to the parental strains B6 and PWD. We detected ~80 instances of NCOs; a reprentative subset of them was validated by Sanger sequencing.

We could detect NCOs as short as below 100bp and as long as above 200bp. The NCOs avoided promoters and overlapped previously published DMC1 (markers of double-strand breaks) and H3K4me3 (markers of PRDM9 binding sites) ChIP-seq peaks from B6, PWD, (B6xPWD)F1 and (PWDxB6)F1. The detected NCOs colocalized with DMC1 hotspots in 86% (PWDxB6, B6xPWD), 75% (B6) and 9% (PWD) of cases. This corresponds well to gradual inbreeding of both *Prdm9* gene and the introgressed PWD chromosomes during the process of formation of consomic strains. It is consistent with an asymmetric behavior of PRDM9 protein when $PRDM9^{B6}$ preferably binds to PWD chromatids and $PRDM9^{PWD}$ preferably binds to B6 chromatids in meiosis. The bases A/T were converted ~3 times ($P<0.0001$) more often to bases G/C than vice versa leading to a genome-wide GC bias.

P-14

Variant Annotation Filtering and Prioritization

Hekel R.[1,3], Budis J.[2,3], Turna J.[3], Szemes T.[1,3]

[1] *Faculty of Natural Sciences of the Commenius University in Bratislava, Slovakia*
[2] *Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia*
[3] *Geneton, s.r.o., Bratislava, Slovakia*

Analyses based on massively paralel sequencing detect vast amounts of variants, while only few are responsible for traits of interest. Successful indentification of these few variants requires annotation with various features, especially function prediction score and conservation score. Different annotation data is scattered across various databases, which makes manual annotation a time-consuming and tedious process.

To facilitate the annotation process, we developed a desktop application called Variant Annotation Analyzer (VAA) together with web based Variant annotation service (VAS). VAA relies on VAS for automatic annotation. The VAS acts as front-end for dbNSFP database which aggregates many of variant databases. We have also developed machine learning prioritization process based on user preference.

VAA supports VCF file format and filtering, sorting, prioritizing and exporting variants to common tabular formats. Modular architecture based on plugins allows easy implementation of new features and additional databases. With the option of further prioritization it provides a powerful tool for fast identification of potential candidate mutations among loads of irrelevant variants. The use of the web service is not limited to the VAA application and is fully open to any academic use.

The presented work was supported by the “REVOGENE – Research centre for molecular genetics” project (ITMS 26240220067) supported by the Operational Programme Research and Development funded by the ERDF.

P-15

Prediction of protein solubility

Hon J.[1,2], Marušiak M.[2], Martínek T.[2], Zendulka J.[2], Bednář D.[1], Damborský J.[1]

[1] *Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic*
[2] *IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic*

Protein solubility is a hallmark for practical use of proteins in biotechnologies and biomedicine. Nowadays, protein solubility poses a major bottleneck in production of many therapeutic and industrially attractive proteins. The key biochemical process limited by protein solubility is a heterologous protein expression – a manipulation of a living organism, usually *E. coli* bacteria, to produce a target protein of interest. Unfortunately, many proteins heterologously expressed in *E. coli* are not sufficiently soluble. Although various experimental approaches aimed at improving protein solubility during heterologous expression were developed, their success is still limited and it is often tedious and expensive to conduct additional experiments. Solubilization attempts are plagued by relatively low success rates and often lead to the loss of biological activity. Therefore, any advance in computational prediction of protein solubility may increase the efficiency and reduce the cost of proteomics studies significantly. Here, we propose a novel protein solubility prediction tool based on the state-of-the-art machine-learning methods. The main contribution lies in improved prefiltering of training dataset and sensible selection of sequence features included in the prediction model. Finally, proper validation and comparison with competing tools was performed using a truly independent test dataset.

P-16

Validation information in the Protein Data Bank: What is it and why should you care?

Smart O. S.[1], Horský V.[2], Gore S.[1], Svobodová Vařeková R.[2], Bendová V.[2,3], Kleywegt G. J.[1], Velankar S.[1]

[1] *Protein Data Bank in Europe, European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, England*
[2] *National Centre for Biomolecular Research, Faculty of Science, and CEITEC – Central European Institute of Technology, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic*
[3] *Institute of Mathematics and Statistics, Masaryk University, Kotlářská 267/2, 611 37 Brno, Czech Republic*

Widespread availability of biomacromolecular structural data has accelerated the progress of research in various life sciences. As an example of this paradigm shift, computer-assisted studies of ligands bound to active sites of proteins and nucleic acids became possible, which in turn aided structure-guided drug discovery and design. Published structures are stored in many databases that have emerged over time, the largest one being the Protein Data Bank (PDB). Experimental methods used by structural biologists steadily improve, which in turn makes steady increase of the number of published structure models per year possible. However, concerns regarding quality of available structures have gone hand-in-hand with broad structure production and usage. Curators of the PDB database have, along with experts from the community of structural biologists, reacted by developing the PDB validation pipeline [1]. It was then integrated into the OneDep structure deposition system, which means that all new structure models are validated during their submission process to the PDB database.

Here, we present the available validation metrics and show how their values can be combined into a single score that can be used to rank macromolecular structures and their domains in search results [2]. This user-friendly score aims to bring validation information closer to the general scientific community, since it does not require extensive experience or knowledge of structural biology on the side of the user.

A major challenge that accompanies crystallographic experiments is how to correctly interpret electron density at binding sites [3]. Such density can represent either ligand molecules, or measured solvent. Incorrect solution of this ambiguity is one of the reasons why quality of ligands in complexes in the PDB is a concerning matter [4]. Therefore, it comes as no surprise that several ligand validation methods are part of the

PDB validation pipeline. Here, we describe these methods. Furthermore, we discuss that the currently used LLDF metric can give misleading results [5].

References

[1] Gore, S. et al. (2017). Structure, 25, 1916–1927.
[2] Smart, O. S., Horský V. et. al. (2018). Acta Crystallographica Section D, 74(3), 237-244.
[3] Smart, O. S. & Bricogne, G. (2015). Multifaceted Roles of Crystallography in Modern Drug Discovery, edited by G. Scapin, D. Patel & E. Arnold, pp. 165–181. Dordrecht: Springer. https://doi.org/10.1007/978-94-017-9719-1_13.
[4] Deller, M. C. & Rupp, B. (2015). J. Comput. Aided Mol. Des. 29, 817–836.
[5] Smart, O. S., Horský V. et. al. (2018). Acta Crystallographica Section D, 74(3), 228-236.

P-17

Variant Calling based on CNN

Hrbek L.[1], Holub J.[1]

[1] *Faculty of Information Technology, Czech Technical University in Prague*

This study deals with *variant calling* problem from *next-generation sequencing* data. Discovered *variants* (differences from reference genome) are used in studies of population genetics, for sequencing new individuals, for identification of causes of disease and for many other analyses.

In contrast to the traditional approach, which explores statistical properties to formulate metrics used for filtering, we propose more general *machine learning* approach. Inspired by the human decision making progress based on optical pattern recognition, our approach employs the *deep convolutional neural network*. Our goal is not only to detect variants at given site but to classify possible systematic mapping and alignment errors as well.

Preliminary work concentrates on sources of errors in the process preceding the variant calling. All relevant information has to be extracted from common file formats and transformed into a more general format understandable by the neural network. Another problem is to identify all desirable labels and obtain their examples in well-annotated data.

Our effort is to provide sufficiently powerful automatic variant caller in form of PC application for a simple workstation. It brings the possibility to train the tool by the customer to maximize performance for the certain project. The data don't have to be provided to other company, in addition. SW solution includes several preprocessing tools and learner using GPU acceleration.

An intention is to provide a *lightweight* variant calling tool accessible for all and free of charge. It promises a good generalization including (sequencing) *platform independence*. Our variant caller can become part of different and customized pipelines due to compatibility with common free tools.

P-18

NGS-based methylation analysis provides insight into the epigenetic landscape of currently endogenizing mule deer gammaretrovirus

Hron T.[1], Elleder D.[1]

[1] *Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic*

Endogenous retroviruses (ERVs) are genetic elements constituting a significant part of the vertebrate genomes. They are generated when an exogenous virus integrates into the host germ line and, subsequently, becomes inherited by successive generations of a host. They are usually fixed in the host population for millions of years and their sequences are damaged by mutations. However, small portion of ERVs retains intact genetic information and have been recently shown to play a key role in various cellular processes and pathologies. Study of the DNA methylation-dependent transcriptional silencing, the main mechanism of host defence against uncontrolled virus propagation in its genome, is crucial for uncovering the ERV-host interactions. Despite the progression in this field, the involvement of epigenetic regulations in the defense against active ERVs is still poorly understood.

We study ERV present in the mule deer genome, CrERV. This evolutionary young virus is extremely polymorphic in its integrations suggesting an ongoing invasion into the host genome. This makes CrERV a unique model for studying retrovirus endogenization. In our work, we employed next generation bisulfite sequencing strategy to determine the methylation pattern of hundreds of CrERV integrations present in different animals. This method offers new insight into the interactions between host and active ERVs.

P-19

Protein family based 2D Diagrams of Secondary Structure Elements

Hutařová Vařeková I.[1,2,3], Hutař J.[1,2], Midlik A.[1,2], Navrátilová V.[4], Svobodová Vařeková R.[1,2], Berka K.[4]

[1] CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
[2] National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
[3] Faculty of Informatics, Botanická 68a, 602 00 Brno-Bohunice, Czech Republic
[4] Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University, 17. listopadu 1192/12, 771 46 Olomouc, Czech Republic

Secondary structure elements (SSEs) such as α-helices and β-sheets are important part of protein structure. Their positions and distances in protein are often characteristic for proteins within whole protein family. Unfortunately, current methods focused on 2D visualization of SSEs (e.g., PROMOTIF [1], Pro-origami [2]) are based on one input protein only. Moreover, they usually do not consider information about real distances of SSEs and therefore SSEs that are close to each other in 3D might be visualized far from each other in 2D. As a result, even when two proteins from the same family differ only slightly in 3D, their 2D SSE diagrams can be totally different.

For this reason, we focused on development of a methodology for 2D SSE diagrams generation, which is based on structures representing whole protein families. In our approach we use three criteria: The first is to minimize the error of SSEs projection from 3D to 2D. Then we concentrate to highlight the similarities of protein families in each protein diagram. Specifically, we use a “skeleton” concept: For each protein family, we find the most firm and stable SSEs. Their position changes minimally in whole family. This “skeleton” get static position in 2D SSE diagrams for proteins from this protein family. The third criterion was to keep the differences between protein 3D structures and transfer them to 2D SSE diagram (classified e.g. by RMSD number). We tested our approach on all protein families described in CATH and this way we showed its applicability.

References

[1] Hutchinson, E. G., Thornton, J. M. (1996). PROMOTIF—a program to identify and analyze structural motifs in proteins. Protein Science, 5(2), 212-220.

[2] Stivala, A., Wybrow, M., Wirth, A., Whisstock, J. C., & Stuckey, P. J. (2011). Automatic generation of protein structure cartoons with Pro-origami. Bioinformatics, 27(23), 3315-3316.

P-22

Framework for knowledge-based prediction of protein-protein interaction sites

Jelínek J.[1], Hoksza D.[1]

[1] *Univerzita Karlova, Matematicko-fyzikální fakulta, Malostranské nám. 25, 118 00 Praha*

Protein-protein interactions (PPI) are crucial in a wide range of biological processes. Demands of their identification by experimental methods lead to development of computational methods. However, their prediction quality is still far from perfect, thus we recently proposed INSPiRE – a new knowledge-based method for prediction of protein-protein interaction sites [1]. Another problem with available methods is that they are typically available as a web server only. As a result, there is not possible a batch processing or it is limited by restrictions of individual web servers, and an integration of these methods into a pipeline is problematic. Also, there are no possibilities to modify individual methods and with only limited possibilities to parametrize them. To avoid these deficiencies, we decided to make our method available as an open-source software.

The software is written in C++ programming language to make it compilable on wide range of platforms and to reach a maximal efficiency. Its back end is designed as a header-only to make its use in other people's projects as easy as possible. The back end is fragmented into independent modules that provide elementary operations (e.g. one module parses temperature of an amino acid, while another module computes its RASA value), so it is easy to write a new module (e.g. generating a new feature) and hook it up. The software also has a front end that allows the user to promptly use INSPiRE. It allows the user to compile the whole pipeline into a single program, or compile each module as a separate program, which allows reusing of common features and/ or knowledge-bases and thus decrease time and space requirements during tuning of the pipeline.

In addition to the prediction of PPI, it is also possible to use the framework to distinguish between biological protein-protein interactions and crystal contacts, or to search homologous proteins or fragments.

References

[1] Jelínek, J., Škoda, P., & Hoksza, D. (2017). Utilizing knowledge base of amino acids structural neighborhoods to predict protein-protein interaction sites. *BMC bioinformatics*, 18(15), 492.

P-23

Real-time impedance based cell assays in bioactivity screening

Kahle M.[1], Bartůněk P.[1]

[1] *Institute of Molecular Genetics of the ASCR, v. v. i.*

Screening of small molecule compounds for bioactivity requires the interaction of these compounds with some biological system; mostly cell lines grown in culture are used. The goal is to detect an effect on as wide selection of targets and pathways as possible.

The effect of such perturbation can be measured among others by gene expression analysis, proteomic analysis, morphological analysis by high-content microscopy or differential cell line viability assays. These methods typically use endpoint measurements but the response of cells varies strongly in time and there is no single time after exposure when all effects can be conveniently measured.

Label-free methods allow to measure properties of cells noninvasively and in real time. They measure the changes in the amount of cells as well as their size, morphology and attachment to the surface. We have validated the impedance based RTCA technique and used it to measure the responses to a library of 2816 bioactive compounds. We will discuss the analysis, visualization and insights gained from this screen.

P-24

Detection of distinct changes in gene-expression profiles in specimens of tumours and transition zones of tenascin-positive/-negative head and neck squamous cell carcinoma

Kolář M.[7], Živicová V.[1,2,3], Gál P.[4,5], Mifková A.[1,2,3], Novák Š.[1,2,3], Kaltner H.[6], Strnad H.[7], Novotný J.[7], Šáchová J.[7], Hradilová M.[7], Chovanec M.[8], Gabius H. J.[6], Smetana K. Jr.[1,9], Fík Z.[1,2,3]

[1] *Charles Univ Prague, Inst Anat, Fac Med 1, Prague, Czech Republic*
[2] *Charles Univ Prague, Fac Med 1, Dept Otorhinolaryngol Head & Neck Surg, Prague, Czech Republic*
[3] *Univ Hosp Motol, Prague, Czech Republic*
[4] *East Slovak Inst Cardiovasc Dis Inc, Dept Biomed Res, Košice, Slovakia*
[5] *Pavol Jozef Šafárik Univ Košice, Fac Med, Dept Pharmacol, Košice, Slovakia*
[6] *Ludwig Maximilians Univ München, Fac Vet Med, Inst Physiol Chem, Munich, Germany*
[7] *Acad Sci Czech Republ, Inst Mol Genet, Lab Genom & Bioinformat, Prague, Czech Republic*
[8] *Charles Univ Prague, Univ Hosp Královské Vinohrady, Dept Otorhinolaryngol & Head & Neck Surg, Fac Med 3, Prague, Czech Republic*
[9] *Charles Univ Prague, BIOCEV, Fac Med 1, Vestec, Czech Republic*

Having previously initiated genome-wide expression profiling in head and neck squamous cell carcinoma (HNSCC) for regions of the tumour, the margin of surgical resecate (MSR) and normal mucosa (NM), we here proceed with respective analysis of cases after stratification according to the expression status of tenascin (Ten).

Tissue specimens of each anatomical site were analysed by immunofluorescent detection of Ten, fibronectin (Fn) and galectin-1 (Gal-1) as well as by microarrays.

Histopathological examination demonstrated that Ten+Fn+Gal-1+ co-expression occurs more frequently in samples of HNSCC (55%) than in NM (9%; $p < 0.01$). Contrary, the Ten−Fn+Gal-1− (45%) and Ten−Fn−Gal-1− (39%) status occurred with significantly ($p < 0.01$) higher frequency than in HNSCC (3% and 4%, respectively). In MSRs, different immunophenotypes were distributed rather equally (Ten+Fn+Gal-1+ ~ 24%; Ten−Fn+Gal-1− ~ 36%; Ten−Fn−Gal-1− ~ 33%), differing to the results in tumours ($p < 0.05$). Absence/presence of Ten was used for stratification of patients into cohorts without a difference in prognosis, to comparatively examine gene-activity signatures. Microarray analysis revealed *i)* expression of several tumour progression-associated genes in Ten+ HNSCC tumours and *ii)* a strong up-regulation of gene expression

assigned to lipid metabolism in MSRs of Ten− tumours, while NM profiles remained similar.

The presented data reveal marked and specific changes in tumours and MSR specimens of HNSCC without a separation based on prognosis.

P-25

Towards universal platform for protein binding site prediction

Krivák R.[1], Jendele L.[1], Novotný M.[2], Hoksza D.[1]

[1] *Charles University, FMP, Department of Software Engineering*
[2] *Charles University, FS, Department of Cell Biology*

Most of proteins perform their function by binding to other molecules – ligands, nucleic acids, peptides, other proteins, etc. In situations, when some type of binding is suspected but only unannotated structure is known, binding site prediction methods can provide useful starting point for further analysis. Existing prediction methods are based either on template matching in a library of known protein complexes (template-based methods) or are template-free (geometric, energetic, conservation-based and machine learning based). Some machine learning based methods for individual types of binding sites exist, but not much attention has been paid to the fact that any of those methods can potentially be used to predict any type of binding sites just by training a model on a different training dataset.

We have created a framework for developing machine learning based binding site prediction methods for various types of binding partners. Resulting methods work by predicting binding score of points lying on the solvent accessible surface of a protein. Those points are described by a feature vector calculated from their local neighbourhood and represent potential locations of atoms of potential binding partners. The system is easily extensible by new protein surface descriptors and integrates a Bayesian optimization procedure for joint optimization of various arbitrary parameters of the algorithm (thresholds, distance cut-offs, etc.). These features allow tailoring developed prediction methods to specific types of binding partners.

So far, we have applied the approach to develop protein-ligand (small molecules) and protein-peptide binding site prediction methods. In both cases, we were able to develop predictors that achieve state-of-the-art performance, while being faster than most of the competing methods.

P-26

Analyzing holobiontic association between host and microbiota in passerines

Kubovčiak J.[1], Kropáčková L.[1], Albrecht T.[1,2], Těšický M.[1], Martin J. F.[3], Kreisinger J.[1,2]

[1] *Department of Zoology, Faculty of Science, Charles University in Prague, Czech Republic*
[2] *Institute of Vertebrate Biology, Academy of Sciences of the Czech Republic, Brno, Czech Republic*
[3] *Centre de Biologie pour la Gestion des Populations, Montferrier-sur-Lez cedex, France*

Increasing number of studies provide evidence of the fundamental influence of gut microbiota (GM) on the physiology, behavior, fitness and evolution of host species. In this context, multicellular organisms are considered as holobionts, ie naturally indivisible entities including the host organism and its associated microbiota. The aim of this study is to analyze GM of wildlife passerines in the Czech Republic and test whether a coevolution process called "phylosymbiosis" takes place between the hosts and their gut microbiota. This phenomenon is a prerequisite to validate the Holobiont concept. Condition of the existence of phylosymbiosis between hosts and their microbiota, is the fact that phylogenetically related species show more similar composition of microbiota. The partial aim is to identify individual components of GM, whose composition is host-species specific and is potentially shaped by coevolutionary process. Using the Illumina MiSeq platform, 164 rRNA bacterial gene amplicons derived from faecal samples of 486 individuals representing 57 species from the Czech Republic were sequenced. OTUs representing the bacterial groups present in GM were generated by DADA2 algorithm, using the expectation-maximization algorithm to correct sequencing errors. The evidence of phylosymbiosis was performed by the Procrustean analysis method along with variance analysis for multidimensional data based on the similarity of samples in the presence of GM bacterial units at different levels of sequence similarity for OTU clustering. To determine this phenomenon for individual GM components, OTUs clustering on a 95% sequence similarity threshold were similarly analyzed. Significant positive dependence of GM composition on the host species included in the study was observed. The progression of this dependence along a scale of sequence similarities thresholds used for individual OTUs definition corresponds to the assumption of the existence of a phylosymbiotic relationship between GM and the host, although this effect is rather mild. Furthermore, individual OTUs, whose composition exhibits a high degree of dependence on the taxonomic relationships of the hosts, have been identified.

P-27

Research projects focused Laboratory Information Management System

Lichvar M.[1], Hekel R.[1,2], Budis J.[1,2,3], Smolak D.[1,4], Radvanszky J.[1,4,5], Szemes T.[1,4,6]

[1] *Geneton Ltd., Bratislava, Slovakia*
[2] *Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia*
[3] *Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*
[4] *Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia*
[5] *Institute for Clinical and Translational Research, Biomedical Research Centre, Slovak Academy of Sciences, Bratislava, Slovakia*
[6] *Comenius University Science Park, Bratislava, Slovakia*

Multi-project research-focused laboratories like core facilities may face challenge to register, track, integrate and monitor samples and their analyses. These data contain personal information, so secure storage and access has high priority, also due to new European legislation. To face these challenges, we have developed a web-based Laboratory Information Management System (LIMS) optimized for multi-project laboratories.

LIMS is web-based application for facilitate storage and bioinformatic analysis of genomic data with emphasis on security of genomic and personal information. Additional computational cluster provides bioinformatic analyses over anonymized genomic data without personal identifiers. Summary analysis reports are immutable to change, but with the possibility of repeated reanalysis in case of upgrades of computational pipelines, as the system is easily extensible over new types and versions of data analysis pipelines.

One instance of LIMS can be used to manage several independent projects, with ability to assign user roles for each project separately. This restricts user to access only those projects, samples and analysis, which they are assigned to. Thus, we can collect a large genomic data sets in a single system, which allows to perform wide-scale population studies that are based on aggregation analyses over large cohort of individuals.

LIMS has an implementation of suggested diseases based on phenotypes terms from HPO database, which may further assist physicians in decision process.

P-28

Genomic single rule learning with an ontology-based refinement operator

Malinka F.[1], Železný F.[1], Kléma J.[1]

[1] *Czech Technical University in Prague, Faculty of Electrical Engineering*

Rule learning is a kind of machine learning method that induces a set of classification rules from a given set of training examples. As a well-known representative of this learners, we can adduce CN2, RIPPER, or PRIM. All of them use if-then statement for corresponding hypothesis formulation where the antecedent is in the form of a conjunction of logical terms, and the consequent is a class label. From a bioinformatician point of view, these learners are suitable especially for their easy and clear interpretation of hypothesis on the contrary of a neural network, for example. The other thing that can help biologists interpret their data in a more natural way is a background knowledge. Nowadays, the most popular form of background knowledge in the field of bioinformatics are ontologies, especially Gene Ontology or Disease Ontology. There are other types of structured databases such as KEGG, that can also be interpreted as an ontology or a taxonomy. In our work, we combine these two concepts, rule learning and ontologies/taxonomies, together.

We propose a new rule learning algorithm that builds classifiers from genomic measurements where the individual entities (e.g. genes and samples) can be structured in a taxonomy or ontology. In particular, we introduce a new refinement operator that with the given ontology significantly reduces a searching space of rules and consequently reduces run time of rule learner in comparison with the traditional refinement operator without a loss of accuracy. The proposed ontology-based refinement operator uses two reduction procedures: a Redundant Generalization that omits candidate rules based on a relation generalization-specialization and a Redundant Non-potential that omits the candidate rules which cannot improve classification accuracy. We demonstrate effectiveness and efficiency of our algorithm on three real genomic datasets.

In future, we plan to extend a form of hypothesis to inducing a rule set instead of a single rule and add negative terms to the antecedent in the rules. Also, we plan to integrate the proposed rule-learning method into our existing semantic biclustering workflow, since it can dramatically reduce its runtime.

This work was supported by the Grant Agency of the Czech Technical University in Prague, grant No. SGS17/189/OHK3/3T/13 and by the Ministry of Health of the Czech Republic, grant No. 17-31398A.

P-31

HERVd update 2018

Pačes J.[1], Moravčík O.[1]

[1] *Ústav molekulární genetiky AV ČR*

Endogenous retroviruses (ERVs) are present in the genomes of practically all vertebrates, as a consequence of infiltration of the host germline lineages by circulating exogenous viruses. A typical mammalian genome contains tens to hundreds of thousands of ERV elements, most of which are evolutionarily old and sustained multiple mutation, deletions and rearrangements. Important roles both in physiology and disease processes have been described for some ERV elements, including regulation of host genes, taking part in placenta formation, and influencing immune responses.

Human endogenous retroviruses (HERVs) cover approximately 10% of the human genome sequence. Various groups of HERVs have been described, ranging in numbers from one to many thousand copies. The pattern of HERV expression has been linked to various diseases, including multiple sclerosis, amyotrophic lateral sclerosis, rheumatoid arthritis and various cancers. Because of the high complexity of HERVs and difficulty in their classification and nomenclature, it is important to provide the scientific community with a database resource of these genetic elements.

This database is compiled from the human genome nucleotide sequences obtained mostly in the Human Genome Projects. We created a relatively simple and fast environment for screening human genome for HERVs. This makes it possible to continuously improve classification and characterization of retroviral families. This is the first publicly available HERV database, that allows users to access individual reconstructed HERV elements, including their sequence, structure and other features. Running from 2002, we were cited more than 70x.

You can search by HERV families, chromosome positions and several other features. Results are linked to other bioinformatics databases, namely genome browsers ENSEMBL and UCSC. Metodology of transposone reconstruction can be used on other genomes as well.

P-32

ScreenX – integrated platform for collection and analysis of chemical compounds and HTS data

Müller T.[1], Jindřich J.[1], Škuta C.[1,2], Sedlák D.[1], Svozil D.[1,2], Bartůněk P.[1]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the ASCR, v.v.i, Vídeňská 1083, 142 20 Prague 4
[2] UCT Prague, Laboratory of Informatics and Chemistry, Technická 5, 166 28 Prague 6

ScreenX is a web-based database/LIMS. It is written mostly in Python programming language and relies on other free and open-source projects like web/database framework Django, relational database system PostgreSQL for data management and cheminformatics framework RDKit for manipulation with chemical structures (chemical format conversion, chemical properties calculation, substructure and similarity search, etc.). GUI uses JavaScript with jQuery and jQueryUI libraries and HTML 5 features, therefore web browser must be compatible.

Following functions were implemented so far: storing information about chemical samples, their properties and structures, substructure and similarity searching, organizing samples into chemical plates, handling of 1D and 2D barcodes for samples and plates, plate reformat (clone, serial dilution, Z-reformat), creating assay plates and arranging them into biological experiments, automatic import of result data to experiments, experiment visualisation and analysis (dose-response curves computation, fragment analysis, cluster analysis, etc.).

Furthermore, core functions of ScreenX will be used for building the European Chemical Biology Database (ECBD), which will be hosted and developed by CZ-OPENSCREEN. This database will contain result data from HTS experiments of EU-OPENSCREEN partner sites and provide them to the community through a user-friendly web interface with the possibility of their further analysis.

P-34

RepeatExplorer: Galaxy Server for In-Depth Characterization of Repetitive Sequences in Next-Generation Sequencing Data

Novák P.[1], Neumann P., Hoštáková N., Macas J.

[1] *Biology Centre CAS*

Repetitive DNA makes up large portions of plant and animal nuclear genomes, yet it remains the least-characterized genome component in most species studied so far. Recent availability of high-throughput sequencing data together with novel bioinformatics tools provide necessary resources for in-depth investigation of genomic repeats and enable large-scale repeat analysis to be run by biologically oriented researchers.

Here we present RepeatExplorer Galaxy server (https://repeatexplorer-elixir.cerit-sc.cz), a collection of software tools for in-depth characterization of repetitive elements, which is accessible via web interface. A key component of the server is the computational pipeline using a graph-based sequence clustering algorithm to facilitate *de novo* repeat identification without the need for reference databases of known elements. Since its first release, number of tools for repeat analysis available on public server has grown up. Today RepeatExplorer include several tools which can be used on both short unassembled NGS reads and genome assemblies. New tools include automatic classification of repetitive sequences based on comprehensive database of transposable element protein domains, genome annotation and classification of repetitive elements, and improved identification of tandem repeats using TAREAN pipeline.

Funding: The work was supported from ERDF/ESF project ELIXIR-CZ: Capacity building (No. CZ.02.1.01/0.0/0.0/16_013/0001777)

References

- Novak, P., Avila Robledillo, L., Koblizkova, A., Vrbova, I., Neumann, P., Macas, J. (2017) – TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. Nucleic Acids Res. 45(12): e111.
- Novak, P., Neumann, P., Pech, J., Steinhaisl, J., Macas, J. (2013) – RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next generation sequence reads. Bioinformatics 29: 792-793.

P-43

Empirical methods for calculation of partial atomic charges – applicability for proteins?

Raček T.[1,2,3], Schindler O.[2], Svobodová Vařeková R.[1,2], Koča J.[1,2]

[1] *CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic*
[2] *National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic*
[3] *Faculty of Informatics, Botanická 68a, 602 00 Brno-Bohunice, Czech Republic*

A concept of partial atomic charges is beneficial in physical and organic chemistry, it was successfully applied in chemoinformatics (e.g., [1]) and its utilization in structural biology and bioinformatics is in progress (e.g., [2]). Conformationally dependent partial atomic charges (i.e., charges considering a conformation of a molecule) can be calculated efficiently via empirical charge calculation methods. A challenging part of the empirical methods is their parameterization, especially parameterization for large molecules like proteins.

We implemented two most popular empirical methods (QEq and EEM) and an extension of one of them (SFKEEM). In parallel, we established a universal parameterization protocol covering all these methods. Afterward, we compared applicability of these methods for datasets containing organic molecules, proteins, and metalloproteins. We found that all the approaches perform well for organic molecules but in the case of proteins and metalloproteins, QEq is the most successful. This method can be a way to calculate partial atomic charges for all available proteins.

References

[1] Svobodová Vařeková, R., Geidl, S., Ionescu, C. M., Skřehota, O., Bouchal, T., Sehnal, D., ..., Koča, J. (2013). Predicting pKa values from EEM atomic charges. Journal of cheminformatics, 5(1), 18.

[2] Ionescu, C. M., Svobodová Vařeková, R., Prehn, J. H., Huber, H. J., Koča, J. (2012). Charge profile analysis reveals that activation of pro-apoptotic regulators Bax and Bak relies on charge transfer mediated allosteric regulation. PLoS computational biology, 8(6), e1002565.

Poster Session

Tuesday, 12. June

Poster session is sponsored by the
National Infrastructure of Chemical Biology and
Protein Data Bank in Europe.

P-21

How to predict structure of fused protein chimeras correctly?

Jarosilová K.[1,2], Vymětal J.[1], Vondrášek J.[1]

[1] *Institute of Organic Chemistry and Biochemistry of the CAS, Flemingovo náměstí 542/2, 166 10 Praha 6*
[2] *Faculty of Science, Charles University, Albertov 6, 128 43 Praha 2*

Allosteric modulation of proteins is a very well known and studied phenomenon. However, do we pay enough attention to the effect of neighboring domains inside the protein structure? Is it possible that inter-domain communication determines specific domain arrangement throughout evolution? And most importantly, can we utilize such features in our favor? We search for an ideal prediction tool that will allow us to predict the structure of fusion proteins correctly, so that the model can be subsequently used either as a prior screen by experimentalists, or as a viable model for further computational studies.

For example, looking at the protein assembly, several types of domains can be found. Some are only observed in one-domain proteins, other can be found only in large multi-domain proteins. Some domains have very strictly defined neighbors, other occur in many different contexts. Remarkably, there are widespread domains that exhibit different properties and functions in different structural context, although their overall 3D structure is the same. Moreover, it seems that this property is common through the proteome of a single organism as well as through the whole phylogenetic tree. We aim to determine what are the best tools to predict and classify such behavior.

To choose an appropriate tool, we have performed a systematic study of artificial fusion of well characterized protein domains. We have created a set of domain combinations that cannot be found in the nature, so that the influence of natural contacts and coevolution between domains would be eliminated. The primary domain of interest is the PDZ domain, which matches the required properties for following reasons: it occurs in plethora of proteins involved in membrane to cytoskeleton signaling, appears both in single and multiple copies in one protein, recent research suggests that its substrate binding is affected by surrounding structural elements and, in various proteins, it displays different specificities and affinities towards its binding partners despite the conservation of its overall tertiary structure. The domain combinations in the set are artificial fusion two-domain chimeras, where the invariable domain is PDZ. The domains in the variable position are artificially prepared domains – rather small, well defined, and reasonably fast folding. On these proteins, we perform both the computational modeling and experimental determination of the structure, to measure the feasibility of different computational prediction models.

We show evidence that the design of new protein domain-combinations is affected by the often-overlooked inter-domain communication. In the future, we aim to apply the results to allosteric modulation mechanism involved in inter-domain communication, with potential pharmacological outcomes.

P-29

Interpretation of QSAR models: mining structural patterns taking into account molecular context

Matveieva M.[1], Cronin M. T. D.[2], Polishchuk P.[1]

[1] *Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Olomouc, Czech Republic*
[2] *School of Pharmacy and Chemistry, Liverpool John Moores University, Liverpool, UK*

The study focused on QSAR model interpretation. Usually QSAR models created by modern machine learning methods are “black boxes” providing no means to interpret the molecular features responsible for the property being studied. The goal of this investigation was to develop a workflow to identify molecular fragments in different contexts important for the property modelled. Using an approach developed earlier – Structural and physicochemical interpretation of QSAR models (SPCI) – we calculated the contributions of fragments and so characterized their relative influence on the properties of a compound. Analysis of the distributions of those contributions using Gaussian mixture modeling (GMM) was performed to identify groups of compounds (clusters) comprising the same fragment, where these fragments had substantially different contributions to the studied property. To detect patterns discriminating those groups of compounds from each other we used SMARTSminer and visual inspection. The approach was applied to analyse the toxicity of 1984 compounds to *Tetrahymena pyriformis*. The results showed that the GMM correctly identified known toxicophoric patterns: it detects groups of compounds where fragments have a specific molecular / mechanistic context making them more “toxic”. This demonstrates the applicability of the this method to interpret QSAR models and retrieve comprehensible and rational patterns, even from data sets consisting of compounds having different mechanisms of action, the analysis of which is difficult to achieve using conventional pattern/data mining approaches.

P-30

Automated Annotation of Secondary Structure Elements for Entire Protein Families

Midlik A.[1,2], Hutařová Vařeková I.[1,2,3], Hutař J.[1,2], Moturu T. R.[1], Navrátilová V.[4], Svobodová Vařeková R.[1,2], Koča J.[1,2], Berka K.[4]

[1] CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno
[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno
[3] Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno
[4] Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University, 17. listopadu 1192/12, 771 46 Olomouc

Protein structural data, deposited in Protein Data Bank, represent a highly valuable source of information and their amount is continuously growing. Currently, the data contain more and more structurally and functionally similar proteins (so called protein families), which originate from various organisms, contain different ligands, or have various mutations. To analyse and examine these data, we must identify comparable and related regions in different proteins from one protein family. A part of this process is annotation of protein secondary structure elements (SSEs), which form stable parts of the protein and help us to localize the key regions.

Since an automated procedure for SSEs annotation is not available yet, we developed such an approach [1]. Our method is template-based, meaning that an annotation of a template protein from the family is provided as an input to the algorithm together with a set of query proteins (the whole family). A three-step algorithm is then executed on each query protein. In the first step, SSEs are detected in the query protein. The second step is structural alignment and superimposition of the query protein and the template protein, so the corresponding parts of the two proteins are located close to each other. The third step is the selection of those SSEs from the query protein which are the best counterparts for the SSEs in the template protein.

References

[1] Svobodová Vařeková, R., Midlik, A., Hutařová Vařeková, I., Hutař, J., Navrátilová, V., Koča, J., Berka, K. (2018). Secondary Structure Elements-Annotations and Schematic 2D Visualizations Stable for Individual Protein Families. Biophysical Journal, 114(3), 46a-47a.

P-33

MOLEonline and ChannelsDB - Tools and Database for Analysis of Biomacromolecular Channels, Tunnels and Pores

Navrátilová V.[1], Berka K.[1], Pravda L.[2,3], Sehnal D.[2,3], Bazgier V.[1], Toušek D.[1,2], Svobodová Vařeková R.[2], Koča J.[2], Otyepka M.[1]

[1] *Dpt Physical Chemistry, RCPTM, Palacký University Olomouc, CZ*
[2] *NCBR, CEITEC, Masaryk University Brno, CZ*
[3] *PDBe, EMBL-EBI, Hinxton, UK*

MOLEonline [1] and ChannelsDB database [2] are interactive, web-based tools for detection and analysis of channels, tunnels and pores in biomacromolecular structures. The updated version of MOLEonline offers easier, simple and fully interactive visualization provided by recently developed LiteMol Viewer which overcomes limitations of the previous version. The application provides two basic modes of calculation: i) computation of the tunnels leading to the specified site within the macromolecule and ii) calculation of transmembrane pores which is available as an automatic pore detection mode. MOLEonline application can use both now obsolete PDB and standard PDBx/mmCIF format and enables analysis of the wide range of biomolecular structures. The tool also brings the connection with other bioinformatics resources – PDBe, OPM, UniProt, CSA and recently developed ChannelsDB database. ChannelsDB contains both channels with the literature reference or channels leading to the biologically important sites (cofactors, active sites) within structures deposited in the Protein Data Bank in Europe [3].

Both MOLEonline and ChannelsDB offer unique analytics for the identification and characterization of channels and pores as well as their geometrical and physico-chemical properties. All is provided free of charge online via internet webpages https://mole.upol.cz and http://ncbr.muni.cz/ChannelsDB/.

References

[1] Pravda L, et al.: MOLEonline: a web-based tool for analyzing channels, tunnels and pores (2018 update). *Nucleic Acids Res*, gky309, 2018. https://mole.upol.cz/
[2] Pravda L, et al.: ChannelsDB: database of biomacromolecular tunnels and pores. *Nucleic Acids Res*, 46(D1), D399–D405, 2018. http://ncbr.muni.cz/ChannelsDB/
[3] Velankar, S. et al.: PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res*, 44, D385-D395, 2016. http://www.ebi.ac.uk/pdbe

P-35

MolMiner: an open-source tool and library for chemical entity extraction

Novotný J.[1,2], Svozil D.[2,3]

[1] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic*
[2] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic*
[3] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic.*

Nowadays, there is great interest from both research and the pharmaceutical industry in efficient access to information about chemical compounds. Unfortunately, no unified system exists for the notation of chemical compounds in the literature. This means that many notation forms (so-called 'chemical entities') are used, including 2D structures that appear as images and, thus, cannot be directly searched for. Therefore, to search effectively for compounds in the literature and further analyse them, we must first extract the chemical entities. This involves identifying them, subsequently converting them into a computer-readable format (e.g. SMILES or InChI) and, in the best case, annotating them in a chemical database. Given this time-consuming process, it is surprising that no one has yet developed an open-source tool able to provide a fully automated extraction, in which the input is a document (e.g. PDF) and the output is a list of extracted chemical entities, especially including 2D structures. And so we present MolMiner, an open-source tool and Python library providing the aforementioned automated extraction. To do so, MolMiner integrates several state-of-the-art open-source tools from the field of chemical information retrieval: OSRA, ChemSpot and OPSIN. MolMiner source code and documentation is freely available on the GitHub website and an end-user package including external software is hosted in the Anaconda Cloud.

P-36

Genomic evolution of *Burkholderia cenocepacia* ST32 during a decade of chronic CF infection

Nunvar J.[1], Capek V.[2], Fiser K.[3], Drevinek P.[1]

[1] *Department of Medical Microbiology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic*
[2] *Bioinformatics Centre, 2nd Faculty of Medicine, Charles University, Prague, Czech Republic*
[3] *Department of Paediatric Haematology and Oncology, 2nd Faculty of Medicine, Charles University and Motol University Hospital, Prague, Czech Republic*

Burkholderia cenocepacia causes severe pulmonary infections in cystic fibrosis patients, which persist for years and might develop into fatal septic pneumonia (cepacia syndrome). To devise new strategies to combat this microbe, it is essential to obtain additional knowledge about the pathogenesis of infection. We conducted a comprehensive genomic analysis of 32 Czech isolates of epidemic clone *B. cenocepacia* ST32, originating from 8 patients.

The analysis identified genes undergoing parallel evolution in multiple patients. Functional predictions for encoded proteins pointed to transition metal metabolism and oxidative stress protection as the key pathways under adaptive evolution during ST32 chronic infection. Surprisingly, the mutations led to attenuation of function; mutations in catalase KatG resulted in impaired detoxification of hydrogen peroxide. Deep sequencing revealed substantial polymorphism in genes of both categories within the pulmonary *B. cenocepacia* ST32 populations, providing independent evidence for selective advantage conferred by mutations in these pathways. In addition, we identified genomic phenomena characteristic for bacteria evolving into host-dependent pathogens (transposition burst of IS elements, genome reduction).

We conclude that in *B. cenocepacia* ST32 chronic infections, there is rapid adaptive evolution which apparently affects a different set of genes than in related species *B. dolosa* (where analogous study was performed). Since the pathways affected by adaptive evolution play role in defence against antimicrobial compounds utilized by immune cells, host immunity appears to shape the progress of *B. cenocepacia* chronic infection.

The work was supported by Ministry of Health, grant No. 15-28017A.

P-37

microRNA – are you real?!

Oppelt J.[1,2], Trachtová K.[1,2], Mráz M.[1,3]

[1] *CEITEC-Central European Institute of Technology, Masaryk University, Brno, Czech Republic*
[2] *National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic*
[3] *Department of Internal Medicine, Hematology and Oncology, University Hospital Brno and Faculty of Medicine MU, Brno, Czech Republic*

MicroRNAs (miRNAs) are together with PIWI-interacting RNAs (piRNAs) and small interfering RNAs (siRNAs) one of the three main classes of small regulatory RNAs. Their crucial role in the regulation of gene expression has been shown in numerous studies. They are involved in both physiological and pathological processes.

MiRBase, a specialized miRNA database, stores information and annotation for various organisms. The annotation of miRNAs is usually based on homology, predictions and assessment of characteristic properties. MiRNAs are also categorized to miRNA families which consist of miRNAs with very similar sequences but often located at different genomic loci. Due to the small miRNA length (~22bp), the annotation contains a lot of false-positive entries. Such misannotations are often caused by other short RNA types as well as fragments of long RNAs which are mistaken for miRNAs. Also, the annotation comprises of a rough estimate of the real sequence, and this is often not the most abundant one observed in the actual sequencing data.

Analysis of miRNA often starts with alignment directly to the miRBase or a genome. Direct miRBase alignment rises at least three possible sources of bias – a) False assignment of long RNA fragments to the annotated miRNA, b) Other short RNAs cross-mapping identified as miRNAs, c) Swapping of miRNAs from the same family due to allowed mismatches. Genome alignment, in addition to the mentioned issues, has to handle massive multi-mapping of the short reads and other random genomic matches.

We have analyzed miRBase annotated miRNAs and identified possible cross-mapping artefacts caused by other short RNAs as well as long RNAs fragments. Additionally, we have assessed the impact of the allowed number of mismatches and read-counting approach on the expression estimates and misassignment. We have evaluated the bias on both fresh-frozen and FFPE samples. As a result, we highlight possible issues in the analysis as well as recommend optimal sample processing.

P-38

Identification of genomic rearrangements in white blood cells of colorectal cancer patients

Ostašov O.[1,2], Pitule P.[1,2], Thiele J. A.[1], Kuhn P.[3]

[1] Biomedical Center, Faculty of Medicine in Pilsen, Charles University, Alej Svobody 76, 32300 Pilsen, Czech Republic.
[2] Department of Embryology and Histology, Faculty of Medicine in Pilsen, Karlovarská 48, 306 05 Plzeň
[3] The Bridge Institute, Dornsife College of Letters, Arts and Sciences, University of Southern California, 3430 S. Vermont Ave., Los Angeles, CA 90089, USA

Colorectal cancer is at the forefront of incidence and mortality associated with cancer globally as well as in Czech Republic. Currently, treatment is based on a combination of surgery followed by oncological treatment in the case of advanced or metastatic colorectal cancers. Appropriate prognostic and predictive markers are required to select the optimal treatment schedule. However, their use may be complicated by evolutionary changes in tumor tissue during treatment, which will affect the effectiveness of the therapy. Tracking these changes using a classic biopsy is often not possible, and therefore, efforts are increasingly being made to use the so-called "fluid" biopsy of free nucleic acids and circulating tumor cells. In this project, however, we focused on the analysis of genomic aberrations in white blood cells in patients with colorectal carcinoma.

In the study, we analyzed a total of 52 white blood cells from 21 patients with colorectal carcinoma. These white blood cells were collected by the micromanipulator from preparations of the peripheral blood nuclear cell suspensions that were used to detect circulating tumor cells. For individual white blood cells, their DNA was amplified using the SigmaAldrich WGA4 kit, the New England Biolabs NEBNext Ultra library was prepared and sequenced on the HiSeq instrument. By analyzing copy number variations in combination with the C5.0 algorithm, we were able to identify a region whose aberrations in white blood cells appear to be associated with duration of tumor-free survival.

This work has been supported by the Charles University Research Fund (Progres Q39), by Charles University Research Centre program UNCE/MED/006 "University Center of Clinical and Experimental Liver Surgery" and by the National Sustainability Program I (NPU I) Nr. LO1503 provided by the Ministry of Education Youth and Sports of the Czech Republic.

P-40

Compound Management in context of employment of cheminformatic tools in practice

Popr M.[1], Bartůněk P.[1]

[1] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics AS CR, v.v.i., Vídeňská 1083, 142 20, Prague 4, Czech Republic.*

Compound Management (CM) is an inseparable part of the of High Throughput Screening (HTS) research field and other drug discovery processes where handling of large collections (typically more than hundreds of thousands) of small organic molecules is to be assured. CM heavily relies on the use of automated robotic systems in connection with dedicated computational technologies such as Laboratory Information Management Systems (LIMS) and various integrated or standalone cheminformatic tools.

At the CZ-OPENSCREEN we use our in-house developed web-based LIMS SW solution ChemGen DB which provides comprehensive set of functionalities necessary for execution of CM workflows. Typical CM workflow starts with the input/import of the inventory data in the database (DB). For higher number of samples, where manual entry is not possible, the import of data files in CSV format is realized. For conversion of large data files obtained from the suppliers a dedicated KNIME workflows are used. KNIME analytical platform SW is equipped with chemistry plugins such as OpenBabel, RDkit, ChemAxon / Infocom, etc. which can be used for interconversion between different chemical structures' notations. Integrity of the inventory data within the ChemGen DB is assured by using unique identifiers for each DB entry together with physical barcoding of all the labware and samples. ChemGen DB also enables to design routine CM procedures such as reformatting, cherrypicking, diluting, etc. and generate lists of instructions for the automation to perform the procedures. CM works in close cooperation with the SW development team and serves as both the end user and beta tester entity of the ChemGen DB. The user feedback and new function requests are realized via JIRA-based communication environment.

P-41

Does lipid metabolism affect centromeric heterochromatin function in fission yeast?

Tvarůžková J.[1], Vishwanatha A.[1], Převorovský M.[1]

[1] *Laboratory of Microbial Genomics, Department of Cell Biology, Faculty of Science, Charles University, Viničná 7, 128 43 Prague 2*

In humans and fission yeast, centromeric heterochromatin is a specialized chromosomal structure that plays an important role in nuclear division. Compromised centromeric heterochromatin can lead to defective mitosis and chromosome loss, resulting in aneuploidy, a hallmark of cancer cells. Carbon metabolism has been linked to both global and targeted changes in histone acetylation levels and gene expression, however, the functional outcomes of these links are still poorly understood. Our initial experiments in fission yeast indicate that altered lipid metabolism affects the integrity of centromeric heterochromatin and mitotic fidelity. We are trying to elucidate how fatty acid synthesis contributes to the regulation of chromatin structure, maintenance of genome integrity, and proper chromosome segregation. We will present pilot ChIP-seq data supporting the existence of novel interconnections between two important cellular processes with fundamental relevance for carcinogenesis: cellular metabolism and chromatin regulation.

P-42

digIS: automated pipeline for detecting distant and novel insertion sequence elements in prokaryotes

Puterová J.[1], Martínek T.[2]

[1] *Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic*
[2] *Department of Computer Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic*

Insertion sequences (ISs) are short DNA sequences that act as a simple transposable element. They play important role in prokaryotic genome structure and evolution. They can affect gene expression, for example by activating neighboring genes leading to antibiotic resistance. Currently available IS annotation tools depend either on existing genome annotations or databases of known IS elements and are not capable to detect novel IS elements or distant members of individual IS families. Here we propose an enhanced approach based on manually curated pHMMs which tackles these shortcomings.

P-44

A pipeline for ncRNA sequence reconstruction and structure characterization of potential homologs from BLAST output

Schwarz M.[1], Vohradsky J.[1], Panek J.[1]

[1] *Institute of microbiology, Czech Academy of Sciences, Videnska 1083 14220 Prague*

The BLAST algorithm is used by many researches as an exploratory RNA sequence search tool. It is extremely useful, but its output includes basically sequence information only, which is not sufficient namely for characterization of sequence fragments. Thus we have developed a pipeline to identify complete sequences of the fragments, predict secondary structures of the subject sequences, and infer their homology to the query RNA.

The pipeline includes several stages: 1) reconstitution of BLAST hits with anchored Locarna algorithm, 2) inference of homology to the query RNA with RSEARCH algorithm 3) prediction of a secondary structure with Centroid-homfold algorithm.

Our pipeline can be used for characterization of ncRNAs in general by extending information included in the BLAST output. Also, it can be extremely useful when homologs of uncharacterized, e.g. newly identified ncRNAs need to be found, and for which more sophisticated methods of homology search can not be used as they require more information of the RNA in their input that is not available.

P-45

Estradiol Dimer Inhibits Tubulin Polymerization and Microtubule Dynamics

Sedlák D.[1]

[1] *IMG AVCR, Prague*

Microtubule dynamics is one of the major targets for new chemotherapeutic agents. Here, we present the synthesis and biological profiling of steroidal dimers based on estradiol, testosterone and pregnenolone bridged by 2,6-bis(azidomethyl)pyridine between D rings. The biological profiling revealed unique properties of the estradiol dimer including cytotoxic activities on a panel of 11 human cell lines, ability to arrest in the G2/M phase of the cell cycle accompanied with the attenuation of DNA/RNA synthesis. Thorough investigation precluded a genomic mechanism of action and revealed that the estradiol dimer acts at the cytoskeletal level by inhibiting tubulin polymerization. Further studies showed that estradiol dimer, but none of the other structurally related dimeric steroids, inhibited assembly of purified tubulin (IC_{50}, 3.6 μM). The estradiol dimer was more potent than 2-methoxyestradiol, an endogenous metabolite of 17β-estradiol and well-studied microtubule polymerization inhibitor with antitumor effects that was evaluated in clinical trials. Further, it was equipotent to nocodazole (IC_{50}, 1.5 μM), an antimitotic small molecule of natural origin. Both estradiol dimer and nocodazole completely and reversibly depolymerized microtubules in interphase U2OS cells at 2.5 μM concentration. At lower concentrations (50 nM), estradiol dimer decreased the microtubule dynamics and growth life-time and produced comparable effect to nocodazole on the microtubule dynamicity. *In silico* modeling predicted that estradiol dimer binds to the colchicine-binding site in the tubulin dimer. Finally, dimerization of the steroids abolished their ability to induce transactivation by estrogen receptor α and androgen receptors. Although other steroids were reported to interact with microtubules, the estradiol dimer represents a new structural type of steroid inhibitor of tubulin polymerization and microtubule dynamics, bearing antimitotic and cytotoxic activity in cancer cell lines.

P-46

FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity

Šícho M.[1,2], de Bruyn Kops C.[1], Stork C.[1], Svozil D.[2,3], Kirchmair J.[1]

[1] Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Computer Science, Center for Bioinformatics, Hamburg, 20146, Germany
[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic
[3] CZ-OPEN-SCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic

Thorough understanding of xenobiotic metabolism plays a crucial role in the development of novel drugs and other bioactive compounds such as cosmetics or agrochemicals. In particular, the identification of atoms where an enzymatic reaction is initiated, so called sites of metabolism (SoMs), can help researchers uncover likely structural changes that a xenobiotic compound may be susceptible to when exposed to metabolic enzymes *in vivo*. In this work, we report on the development of FAst MEtabolizer 2 (FAME 2) [1], a SoM prediction model and software. The methodology of FAME 2 is very simple and relies on the extra trees machine learning algorithm in combination with a novel set of 2D circular atomic descriptors. This enables FAME 2, which is distributed free of charge from the authors, to make fast regioselectivity predictions that require little to no effort from the user, but reach accuracy comparable to more complex methods.

References

[1] Šícho, M.; de Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. J. Chem. Inf. Model. 2017 DOI: 10.1021/acs.jcim.7b00250.

P-47

New Developments in the Molpher-lib Chemical Space Exploration Library: algorithms, substructure locking and programming interface

Šícho M.[1], Svozil D.[1,2]

[1] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic.
[2] CZ-OPEN-SCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic.

In the past, we reported on the development of Molpher-lib [1], a C++ and Python software library for *in silico* chemical space exploration and focused virtual library design. Molpher-lib uses a unique atom-based approach called molecular morphing. This approach acts on the premise that if structural features of known bioactive compounds are combined to create new structures, the resulting hypothetical molecules will likely have similar activity themselves. Molpher-lib provides a simple and comprehensive programming interface that can be used not only for easy implementation and customization of molecular morphing algorithms, but also lends itself to other endeavors such as sampling chemical space around particular scaffolds. In this work, we present two new exploration algorithms as well as substructure locking that enables the user to lock parts of the molecule against some or all modifications, which can be useful in the determination of the impact that certain structural changes could have on pharmacokinetics and pharmacodynamics of a compound. We also summarize new features of the programming interface itself that now allows, among other things, custom implementations of morphing operators, which we envision as important to extend the capabilities of Molpher-lib even beyond chemical space exploration.

References

[1] Šícho, M.; Svozil, D. Molpher-lib: softwarová knihovna pro systematickou exploraci chemického prostoru. Presented at ENBIK 2016, Loučeň, June 2016.

P-48

Sequence based classification of bacteriophage

Baláž A.[2,4], Smoľak D.[1,4], Budiš J.[2,4], Kajsík M.[3], Böhmer M.[1], Szemes T.[1,3,4]

[1] *Faculty of Natural Sciences; Comenius University; Bratislava*
[2] *Faculty of Mathematics, Physics and Informatics; Comenius University; Bratislava*
[3] *University Science Park; Comenius University; Bratislava*
[4] *Geneton Ltd; Bratislava*

Increased bacteria resistance to antibiotics treatment represents emerging problem in medical care. The promising way is to make use of viruses, that infect bacteria, also called bacteriophages. The main purpose of this research is identification of hosts of particular bacteriophage from its genomic sequence. We based our classification on genes that are specific to certain genera of host.

We downloaded Fasta records from NCBI, Viralzone and Phagedb databases with information about their hosts. Prokka software was using to extract genes from genomic sequences. We created subsets of most common hosts across our dataset. Next step was generating of clusters of similar genes using Markov clustering. Each phage was then represented as a binary vector, where each component indicate if the phage contains a gene from the cluster. Finally, we generated binary classificators for this reduced representation.

We propose the computational pipeline, that produces classification models for prediction of host from phage sequences. We believe that our pipeline will provide deeper insight into molecular mechanisms of phages, mainly their ability to infect specific genera of bacteria.

P-49

Report on Evaluating Quality and Reliability of the Next Generation Sequence Data with a Mock Community

Špánek R.[1,2], Dolinová I.[1,2]

[1] *Institute for Nanomaterials, Advanced Technologies and Innovation, Technical University of Liberec (TUL), Studentská, 46117 Liberec, Czech Republic.*
[2] *Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic*

We report results on analysing of predefined mock community samples using Mothur software and the Ion Torrent sequencing platform (Life Technologies, USA). The mock community was designed by combining several of well-characterized bacterial species. In our experiment which took place over last 3 years we did use mock community consisting of the following bacteria species - Pseudomonas alcaliphila JAB1, Pandoraea pnomenusa B-356, Rhizobium radiobacter C58, Achromobacter xylosoxidans A8, Burkholderia xenovorans LB400, Cupriavidus necator H850, Pseudomonas veronii 20a2, Methylobacterium radiotolerans JCM 2831, Rhodococcus jostii RHA1, Arthrobacter chlorophenolicus A6, Bacillus pumilus SAFR-032, Micrococcus luteus Fleming strain. Processing the mock community with the same pipelines allows to evaluate the quality of the sequencing outputs and also to infer reliability of obtained results. The goal of the experimental was to determine a quality and reliability of IonTorrent platform, reproducibility of our experiments and particularly influence of quality and aging of a chemistry used for preparing amplicon libraries. The obtained outputs show very interesting trends in change of quality and length of reads produced by different chemistry as well as very severe impact of aging of chemistry on length of sequences. The secondary output of the report is experience with Mothur and IonTorent sequencing technology for probing microbial communities in samples coming mainly from polluted ground and underground waters.

P-50

CAVERDOCK: A New Tool for Analysis of Ligand Transport Processes in Proteins Using Molecular Docking

Stourac J.[1,2], Vavra O.[1,2], Filipovic J.[3], Plhak J.[3], Marques S. M.[1,2], Bednar D.[1,2], Matyska L.[3], Damborsky J.[1,2]

[1] Loschmidt Laboratories, Department of Experimental Biology and Research Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, Kamenice 5/A13, 625 00 Brno, Czech Republic
[2] International Centre for Clinical Research, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic
[3] Institute of Computer Science, Faculty of Science, Masaryk University, Botanicka 554/68a, 60200 Brno, Czech Republic

Understanding the mechanisms of ligand binding and unbinding into the proteins' is crucial for drug design and protein engineering. Since many proteins have their active/binding sites buried inside the protein core, properties of access paths connecting the protein surface with the active site can heavily influence the whole binding mechanism [1]. We developed a novel software tool called CaverDock to study the ligand passage through such pathways [2, 3]. The tool is based on the idea of slicing the access tunnels into the small disks followed by iterative molecular docking calculation by AutoDock Vina [5], employing the constraints to every single disk. The software requires protein structure, tunnel topology from CAVER [4] and ligand structure as the inputs. The outputs are continuous ligand trajectory, estimated free energy of binding along the pathway, activation barrier of the transport process and the energy difference between bound and unbound states. CaverDock is easy to setup and very fast - a typical calculation time in dozens of minutes makes it suitable even for a large scale virtual screenings. CaverDock is available free of charge at the website https://loschmidt.chemi.muni.cz/caverdock/.

References

[1] Marques, S.M., et al. 2016: Enzyme Tunnels and Gates as Relevant Targets in Drug Design. *Medicinal Research Reviews* 37: 1095-1139.
[2] Vavra, O., et al. CAVERDOCK: A New Tool for Analysis of Ligand Binding and Unbinding Based on Molecular Docking. In preparation.
[3] Filipovic, J., et al. A Novel Method for Analysis of Ligand Binding and Unbinding Based on Molecular Docking. In preparation.
[4] Chovancova, E., et al. 2012: CAVER 3.0: A Tool for Analysis of Transport Pathways in Dynamic Protein Structures. *PLOS Computational Biology* 8: e1002708.

[5] Trott, O. & Olson, A.J., 2010: AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization and Multithreading. *Journal of Computational Chemistry* 31: 455-461.

P-51

Connection of glycoinformatics databases with Pubchem

Suchánková P.[1,2], Svobodová Vařeková R.[1,2]

[1] CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
[2] National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic

Nowadays, information about structure of organic molecules and proteins is stored in general databases and described via common data formats. Unfortunately, glycoinformatics databases are isolated and use specialised data formats. For better work with interesting data stored in these databases, it is necessary to link together information from these specialized databases with general-purpose database.

Among such specialised databases belong for example GlyConnect and UnicarbKB – databases of glycan structures, which utilise GlycoCT data format [1]. We developed a workflow for translation of GlycoCT format to common molecular formats SMILES, InChI and InChIKey. Based on these standard formats, we established a linkage between GlyConnect, UnicarbKB and PubChem [2] databases. We will present the workflow and information describing a matching between these databases.

References

[1] Herget S, Ranzinger R, Maass K, Lieth CW. GlycoCT—a unifying sequence format for carbohydrates. Carbohydrate Research. 2008, 343(12), 2162-2171.
[2] Sunghwan K, Thiessen PA, Bolton EE, et al. PubChem Substance and Compound databases. Nucleic Acids Research. 2016, 44(D1), D1202-D1213.

P-52

Changes in gene expression during immortalization of haematopoietic cells

Svatoňová P.[1,2], Kolář M.[1], Bartůněk P.[3], Svoboda O.[3], Machoňová O.[3], Oltová J.[3]

[1] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic*
[2] *Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic*
[3] *Laboratory of Cell Differentiation, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic*

Signalling pathways are crucial for correct cell functioning and communication. The pathways respond to various extracellular signalling molecules including growth factors and their disruption can lead to pathological conditions, such as tumor diseases and other malignancies. We studied an Epo/EpoR-dependent haematopoietic cell line from zebrafish (*Danio rerio*) that lost dependence on stem cell factor (SCF) in two consecutive steps of immortalization. In a time-series based experiment, gene expression was analysed on a whole genome scale using RNA-Seq. The resulting data was analysed to obtain genes that changed in transcription level most significantly. Large number of differentially expressed genes indicated that there were several processes, which accounted for observed changes. Thus, we divided the differentially expressed genes in clusters according to time dependence of their expression profiles. We characterized the resulting gene clusters functionally with respect to the Gene Ontology terms and KEGG signalling and metabolic pathways.

P-55

Correction of force field for drug-like molecules using property map collective variable

Trapl D.[1], Spiwok V.[1]

[1] *Department of Biochemistry and Microbiology, UCT Prague*

The accuracy of molecular simulations depends on an empirical molecular mechanics potential known as a force field. While force fields designed for proteins or nucleic acids are considered accurate, force fields for drug-like molecules still need many improvements. Here, we present a novel approach for force field correction tailored to a general drug-like compound. Using property map collective variable, it is possible to approximate a certain conformationally dependent property by a weighted average of this property for a series of representative landmark structures. As a value of property we have chosen the difference between potential energies of selected conformers calculated by accurate (force field) and inaccurate potential (quantum chemical methods). To validate this method we used seven AMBER force fields and we performed a set of 20-ns-long metadynamics simulations of Ace-Ala-Nme in water. We generated 144 landmark structures of Ace-Ala-Nme differing in values of torsion phi and psi. And then we tried to transform one force field (e.g. AMBER94) to another one (e.g. AMBER03). The obtained free energy surfaces of the corrected force fields (e.g. AMBER94 corrected to AMBER03) and the intended ones (e.g. AMBER03 without correction) were in good agreement. Furthermore, we used same landmark structures and differences between potential energy obtained using force field and DFT calculation. We also present force field correction for important anticancer drug Imatinib as a use case example. Our method appears suitable for adjusting force field for general drug-like molecule.

P-56

Into the wild: expression and characterization of random protein libraries

Tretyachenko V.[1], Vymětal J.[2], Bednarová L.[2], Fujishima K.[3], Vondrášek J.[2], Hlouchová K.[1,2]

[1] *Charles University in Prague, Faculty of Science, Department of Biochemistry, Hlavova 2030, 128 00, Praha 2, Czech Republic*
[2] *Institute of Organic Chemistry and Biochemistry, Flemingovo nám. 2, 166 10, Praha 6, Czech Republic*
[3] *Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, 1528550, Japan*

Modern protein science heavily relies and benefits from the data generated from experimental characterization of natural protein sequences. Our study moves beyond the natural world in attempt to construct and describe the behaviour of random protein libraries without any evolutionary background. In order to investigate the structure-forming potential of random proteins we designed and applied a novel random library construction and purification methodology. Our libraries, although being random in sequence, are customized in amino acid content and ratios. This approach allowed us to study the secondary structure content of (i) natural-like random proteins composed of all 20 amino acids, library of proteins built from amino acids present in (ii) prebiotic and (iii) early biotic world and protein library made from (iv) minimal set of amino acids from the rational protein design point of view. In addition to experimental characterization we performed bioinformatic screening of random libraries in order to unveil structural landmarks of different amino acid alphabets.

P-57

SEED 2: a user-friendly tool for amplicon high-throughput sequencing data analyses

Větrovský T.[1], Baldrian P.[1], Morais D.[1]

[1] *Institute of Microbiology of the CAS, Videňska 1083, 14220 Prague 4, Czech Republic*

Modern molecular methods have increased our ability to describe microbial communities. Along with the advances brought by new sequencing technologies, we now require intensive computational resources to make sense of the large numbers of sequences continuously produced. The software developed by the scientific community to address this demand, although very useful, require experience of the command-line environment, extensive training and have steep learning curves, limiting their use. We created SEED 2, a graphical user interface for handling high-throughput amplicon-sequencing data under Windows operating systems. SEED 2 is the only sequence visualizer that empowers users with tools to handle ampliconsequencing data of microbial community markers. It is suitable for any marker genes sequences obtained through Illumina, IonTorrent or Sanger sequencing. SEED 2 allows the user to process raw sequencing data, identify specific taxa, produce OTU-tables, create sequence alignments and construct phylogenetic trees. Standard dual core laptops with 8 GB of RAM can handle ca. 8 million of Illumina PE 300 bp sequences, ca. 4GB of data. SEED 2 was implemented in Object Pascal and uses internal functions and external software for amplicon data processing. SEED 2 is a freeware software, available at http://www.biomed.cas.cz/mbu/lbwrf/seed/ as a self-contained file, including all the dependencies, and does not require installation.

P-58

SYBA: Fragment based prediction of hard-to-synthesize structures

Voršilák M.[1,2], Svozil D.[1,2]

[1] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czech Republic.*
[2] *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Institute of Molecular Genetics, AS CR v.v.i., Prague, Czech Republic.*

Machine learning methods are often used in cheminformatics to predict activity, to cluster similar structures or to classify structures into distinctive classes. While a biological activity or toxicity can be experimentally measured, another important molecular property, the synthetic feasibility, is a more abstract feature that can't be easily assessed. Furthermore, synthetic feasibility is not only abstract, but even hard-to-synthesize (HS) structures are not readily available from any database, which is needed for classification. Thus, we developed HS structure generator called Nonpher to address this issue [1].

Classification based on Bayes probability relies on the sum of feature contributions that are calculated as differences between feature probabilities in training datasets. As a feature space, we utilized fragments generated with Morgan algorithm. Random samples of the ZINC database were analyzed to obtain natural (easy-to-synthesize, ES) fragment scores, conversely, Nonpher generated dataset to obtain HS fragment scores and the combination of individual fragment scores assemblies SYBA classifier. Our model promises easy interpretation because ES and HS fragments are known and on the test set also high accuracy.

References

[1] M. Voršilák, D. Svozil, J. Cheminform. 2017, 9:20.

P-59

Effect of phosphorylation on intrinsic conformational preferences of serine, threonine and tyrosine. A computational study

Vymětal J.[1], Vondrášek J.[1]

[1] *Institute of Organic Chemistry and Biochemistry of CAS, Fleminovo nam. 542/2, 166 10 Praha 6, Czech Republic*

Phosphorylation of serine, threonine and tyrosine ranks among the most frequent and crucial post-translational modifications (PTM) of proteins. Approximately one-third of all proteins in an eucaryotic (human) cell are estimated to be phosphorylated by at least one of 500 putative kinases and de-phosphorylated by 150 phosphatases. A large portion of phosphorylation sites are located in disordered regions, and intrinsically disordered proteins themselves are preferred substrates of protein kinases.

We studied effects of phosphorylation on behavior of a protein chain by molecular dynamics and enhanced sampling methods (metadynamics). The current biomolecular force fields such as amber and charmm contain parameters for phosphorylated serine, threonine and tyrosine enabling simulations of these frequent PTM residues. However, quality of these parameters is the critical point because they have stood aside the extensive force-field testing.

We utilized terminally capped amino acids (dipeptides) as the simplest model molecules for intrinsically disordered regions in proteins. Our simulations revealed a large heterogeneity of the conformational ensembles sampled by all model dipeptides. We analyzed the net effects of phosphorylation on population of backbone conformers and side-chain rotamers, as well as their dependence on the particular force field involved in our study.

We found that phosphorylation induced quantitative changes of different magnitudes in the populations of individual conformers in the simulated ensembles. Indeed, we identified these effects as a mechanism how the net structural trends can be affected by PTM. However, the force fields involved in our study produced rather distinct ensembles and, therefore, non-consistent structural trend. The ability of a force field to describe correctly the ensemble of preferred conformers in quantitative manner seems to be critical for its performance. Nevertheless, even the net structural trends are difficult to extract from experimental measurements and they provide only a indirect guidance for further force field development. Additionally, we discuss other routes for the parametrization and the limits of the currently used functional forms.

Agea Lorente Maria Isabel *(iagea1993@gmail.com)*
Department of Informatics and Chemistry, UCT Prague

Bartoš Oldřich *(124600@seznam.cz)*
ÚŽFG AV ČR, v. v. i., Liběchov

Bazgier Václav *(vaclav.bazgier@upol.cz)*
Katedra fyzikální chemie, Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Berka Karel *(karel.berka@upol.cz)*
Palacky University, Olomouc

Bouška Luděk *(Ludek.Bouska@vscht.cz)*
VŠCHT Praha

Brejová Bronislava *(brejova@dcs.fmph.uniba.sk)*
Fakulta matematiky, fyziky a informatiky, Univerzita Komenského v Bratislave

Budiš Jaroslav *(jaroslav.budis@gmail.com)*
Geneton s.r.o., Bratislava

Cvacho Ondřej *(cvachond@fit.cvut.cz)*
Czech technical university in Prague

Čajka Tomáš *(tomas.cajka@fgu.cas.cz)*
Institute of Physiology CAS, Department of Metabolomics, Prague

Čech Petr *(cechp@vscht.cz)*
VŠCHT Praha

Červeňanský Michal *(misko.cervenansky@gmail.com)*
Institute of the Biophysics of the Czech Academy of Sciences, Brno

Čmelo Ivan *(cmeloi@vscht.cz)*
VŠCHT Praha

Damborský Jiří *(jiri@chemi.muni.cz)*
Masarykova univerzita, Brno

Dehaen Wim *(wimdehaen@gmail.com)*
VŠCHT Praha

Dohnalová Hana *(hannicka12@seznam.cz)*
VŠCHT Praha

Doležalová Marie *(d.marienka@seznam.cz)*
Biologické centrum AV ČR, v. v. i., České Budějovice

Duris Frantisek *(frantisek.duris@geneton.sk)*
Geneton s.r.o., Bratislava

Ehler Edvard *(edvard.ehler@img.cas.cz)*
Institute of Molecular Genetics, ASCR, Praha 4

Fesl Jan *(fesl@post.cz)*
Ústav aplikované informatiky, Jihočeská univerzita, České Budějovice

Galgonek Jakub *(jakub.galgonek@uochb.cas.cz)*
Institute of Organic Chemistry and Biochemistry of the CAS, Praha

Gergelits Václav *(vaclav.gergelits@img.cas.cz)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha 4

Harazim Markéta *(markeeta.sh@gmail.com)*
Institute of Vertebrate Biology, The Czech Academy of Sciences, Brno

Hekel Rastislav *(rtcz.svk@gmail.com)*
Geneton s.r.o., Bratislava

Hon Jiří *(ihon@fit.vutbr.cz)*
Přírodovědecká fakulta, Masarykova Univerzita, Brno

Horský Vladimír *(vladimir.horsky@mail.muni.cz)*
CEITEC MU, Brno

Hoštáková Nina *(nhostakova@gmail.com)*
Biology Centre CAS, České Budějovice

Hrbek Lukáš *(hrbeklu1@fit.cvut.cz)*
České vysoké učení technické v Praze, Praha 6

Hron Tomas *(tomas.hron@img.cas.cz)*
HPST, s. r. o., Praha

Hutařová Vařeková Ivana *(ivarekova@centrum.cz)*
Masaryk University, Brno

Jakubec Dávid *(david.jakubec@uochb.cas.cz)*
Institute of Organic Chemistry and Biochemistry of the CAS, Praha 6

Jansík Branislav *(branislav.jansik@vsb.cz)*
IT4I, VSB-TUO, Ostrava-Poruba

Jarosilová Kateřina *(K.Jarosilova@seznam.cz)*
Institute of Organic Chemistry and Biochemistry of the CAS, Prague

Jedlička Pavel *(jedlicka@ibp.cz)*
Biofyzikální ústav, Akademie věd České republiky, v. v. i., Brno

Jelínek Jan *(jelinek@ksi.mff.cuni.cz)*
Charles University, Faculty of Mathematics and Physics, Praha 2

Kahle Michal *(kahle@img.cas.cz)*
Institute of Molecular Genetics of the ASCR, Prague

Kolář Michal *(kolarmi@img.cas.cz)*
Institute of Molecular Genetics ASCR, Prague 4

Krivák Radoslav *(krivak@ksi.mff.cuni.cz)*
Univerzita Karlova - MFF, Praha 1

Kubovčiak Jan *(jakubovciak@gmail.com)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha

Kumazawa Morais Daniel *(daniel.morais@brmicrobiome.org)*
Institute of Microbiology of the CAS, Prague 4

Lexa Matej *(lexa@fi.muni.cz)*
Masaryk University, Brno

Lichvár Michal *(michal.lichvar@geneton.sk)*
Geneton s.r.o., Bratislava

Macas Jiří *(macas@umbr.cas.cz)*
Biology Centre CAS, České Budějovice

Malinka František *(malinfr1@fel.cvut.cz)*
Czech Technical University in Prague

Martínková Natália *(martinkova@ivb.cz)*
The Czech Academy of Sciences, Institute of Vertebrate Biology, Brno

Matveieva Mariia *(mariia.matveieva@upol.cz)*
Ústav molekulární a translační medicíny Lékařská fakulta Univerzity Palackého v Olomouci

Midlik Adam *(midlik@mail.muni.cz)*
Masaryk University, Brno

Modrák Martin *(martin.modrak@biomed.cas.cz)*
Institute of Microbiology of the Czech Academy of Sciences, Prague

Mokrejš Martin *(martin.mokrejs@vsb.cz)*
IT4Innovations, VŠB – Technical University of Ostrava

Moravčík Ondřej *(ondrej.moravcik@img.cas.cz)*
Ústav molekulární genetiky AV ČR, v. v. i., Praha 4

Musil Miloš *(imusilm@fit.vutbr.cz)*
Masarykova univerzita, Přírodovědecká fakulta, Brno

Müller Tomáš *(tomas.muller@img.cas.cz)*
Institute of Molecular Genetics of the ASCR, v. v. i., Prague

Navrátilová Veronika *(veronika.navrat@gmail.com)*
Palacký University, Olomouc

Novák Petr *(petr@umbr.cas.cz)*
Biology Centre CAS, České Budějovice

Novotný Jiří *(jiri.novotny@img.cas.cz)*
Institute of Molecular Genetics of the ASCR, v. v. i., Prague 4

Ntie-Kang Fidele *(ntiekfidele@gmail.com)*
University of Chemistry and Technology, Prague 6

Nunvář Jaroslav *(botanik@atlas.cz)*
2. lékařská fakulta UK, Praha

Oppelt Jan *(jan.oppelt@mail.muni.cz)*
NCBR & CEITEC MU, Brno

Ostašov Pavel *(pavel.ostasov@lfp.cuni.cz)*
Lékařská fakulta v Plzni, Univerzita Karlova, Plzeň

Pačes Jan *(hpaces@img.cas.cz)*
Institute of Molecular Genetics AS CR, Praha

Panek Josef *(panek@biomed.cas.cz)*
IMIC ASCR, Prague 4

Pitule Pavel *(pavel.pitule@lfp.cuni.cz)*
Lékařská fakulta UK v Plzni

Polishchuk Pavlo *(pavlo.polishchuk@upol.cz)*
Ústav molekulární a translační medicíny, Univerzita Palackého v Olomouci

Popr Martin *(popr@img.cas.cz)*
Institute of Molecular Genetics of the ASCR, v. v. i., Prague

Pravda Lukáš *(lpravda@ebi.ac.uk)*
Protein Data Bank in Europe, EMBL-EBI, Hinxton

Převorovský Martin *(prevorov@natur.cuni.cz)*
Univerzita Karlova, Přírodovědecká fakulta, Praha 2

Puterová Janka *(iputerova@fit.vutbr.cz)*
Fakulta informačních technologií Vysoké učení technické v Brně

Raček Tomáš *(tom@krab1k.net)*
Masarykova univerzita, Nový Jičín

Raus Martin *(martin.raus@upol.cz)*
CRH - Odd.biochemie proteinů a proteomiky, Univerzita Palackého v Olomouci

Reigl Tomáš *(tomas.reigl@gmail.com)*
CEITEC MU, Brno

Rose Alexander *(alexander.rose@weirdbyte.de)*
UC San Diego, San Diego

Roy Sudeep *(roysudeep28@gmail.com)*
Brno University of Technology, Brno

Röslein Jan *(rose.jan@email.cz)*
ÚŽFG AV ČR, v. v. i., Liběchov

Schwarz Marek *(marek.schwarz@biomed.cas.cz)*
Mikrobiologický ústav AV CR, Praha

Sedlák David *(sedlak@img.cas.cz)*
IMG AVCR, Prague

Sehnal David *(david.sehnal@hotmail.com)*
CEITEC Masaryk University, Brno

Smoľak Dávid *(d.smolak@outlook.com)*
Geneton s. r. o., Bratislava

Stočes Štěpán *(stepan.stoces@seqme.eu)*
SEQme s.r.o., Dobříš

Strnad Hynek *(strnad@img.cas.cz)*
Institute of Molecular Genetics of the ASCR, v. v. i., Prague

Suchánková Pavla *(410428@mail.muni.cz)*
Faculty of science Masaryk University, Brno

Sumbalová Lenka *(lenka@sumbal.cz)*
Masarykova univerzita, Přírodovědecká fakulta, Brno

Svatoňová Petra *(svatonovapeta@email.cz)*
Institute of Molecular Genetics of the ASCR, v. v. i., Prague

Svobodová Vařeková Radka *(radka.svobodova@ceitec.muni.cz)*
Masaryk University, Brno

Svozil Daniel *(svozild@vscht.cz)*
UCT Prague

Šícho Martin *(sichom@vscht.cz)*
UCT Prague

Škuta Ctibor *(ctibor.skuta@img.cas.cz)*
CZ-OPENSCREEN, Institute of Molecular Genetics of the ASCR, v. v. i., Prague 4

Špánek Roman *(roman.spanek@tul.cz)*
TU v Liberci

Štourač Jan *(stourac.jan@gmail.com)*
Masarykova Univerzita - Přírodovědecká fakulta, Brno

Trapl Dalibor *(dalibor.trapl@vscht.cz)*
University of Chemistry and Technology, Praha

Treťjačenko Vjačeslav *(tretyacv@natur.cuni.cz)*
Charles University, Prague

Větrovský Tomáš *(kostelecke.uzeniny@seznam.cz)*
Institute of Microbiology CAS, Prague

Vinař Tomáš *(tomas.vinar@fmph.uniba.sk)*
Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Bratislava

Vohradský Jiří *(vohr@biomed.cas.cz)*
Institute of Microbiologu ASCR, v. v. i., Prague

Vondrak Tihana *(vondrak@umbr.cas.cz)*
Biology centre CAS, Ceske Budejovice

Voršilák Milan *(vorsilam@vscht.cz)*
Institute of Molecular Genetics of the ASCR, v. v. i., Prague 4

Vymětal Jiří *(jiri.vymetal@uochb.cas.cz)*
Ústav Organické Chemie a Biochemie AVČR, Praha 6