ENBIK2022 conference proceedings

Editors

Petr Čech, Daniel Svozil

Prague 2022

ENBIK2022 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 978-80-7592-127-7

SESSION 1

Structures

L1-01

Discovering the general architecture of protein families with OverProt

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

[1] CEITEC - Central European Institute of Technology, Masaryk University, Brno, midlik@mail.muni.cz
[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno
[3] Faculty of Informatics, Masaryk University, Brno
[4] Department of Physical Chemistry, Faculty of Science, Palacký University, Olomouc

Every protein family has a set of characteristic secondary structure elements (SSEs) – helices and β-strands. Their number, order, and spatial arrangement is relatively consistent throughout the whole family; thus they define the general architecture of the family and provide a good guide for orientation within the structures. However, there are always some variations within the family, and a single structure is not enough to represent the general architecture of the whole family. Therefore we developed OverProt, which gathers the secondary structure information from all family members and creates the secondary structure consensus. This consensus shows the general architecture of the family as well as its variation, and thus provides a useful insight into the family (just as the sequence logo does for a family of sequences). OverProt server (https://overprot.ncbr.muni.cz/) provides precomputed consensus for all CATH superfamilies plus user-defined computations, visualized by an interactive viewer, which shows the SSE type, length, frequency of occurrence, spatial variability, and β-connectivity. OverProt is also utilized in the visualization tool 2DProts [1], which has been integrated into the CATH database.

References

[1] Hutařová Vařeková I, Hutař J, Midlik A, Horský V, Hladká E, Svobodová R, Berka K (2021) 2DProts: database of family-wide protein secondary structure diagrams. Bioinformatics, 37, 4599–4601.

L1-02

Protein structure quality trends

Svobodová R.[1,2], Horský V.[1,2], Bučeková G.[1,2], Porubská J.[1,2], Doshchenko V.[2]

[1] CEITEC - Central European Institute of Technology, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic

Information about protein 3D structure, collected in the Protein Data Bank (PDB), is a valuable input for structural bioinformatics research and other life science fields. An important topic is the quality of the data because the reliability of research results based on these protein structure data depends on their quality. For this reason, there is a strong focus on protein structure validation. Various quality criteria were published, and information about the quality of individual protein structures is summarised in the PDB validation reports [1].
Here, we utilise our tools ValTrendsDB [2] and ValidatorDB [3] and examine selected interesting trends in protein structure quality. First, we focus on the evolution of protein structure quality in time. Afterwards, we analyse the relationship between the quality of a structure and the journal in which it has been published. Last but not least, we uncover several ligand quality issues, which are not considered by common validation procedures yet.

References

[1] Smart, O. S., Horský, V., Gore, S., Svobodová Vařeková, R., Bendová, V.,Kleywegt, G. J., Velankar, S. (2018). Worldwide Protein Data Bank validation information: usage and trends. Acta Crystallographica Section D, 74, 237-244.
[2] Horský, V., Bendová, V., Toušek, D., Koča, J., Svobodová Vařeková, R. (2019). ValTrendsDB: bringing Protein Data Bank validation information closer to the user. Bioinformatics, 35(24), 5389-5390.
[3] Sehnal, D., Svobodová Vařeková, R., Pravda, L., Ionescu, C.-M., Geidl, S., Horský, V., Jaiswal, D., Wimmerová, M., Koča, J. (2015). ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank. Nucleic acids research, 43(D1), D369-D375.

L1-03

A Structure Validation Concept Beyond the Static Resolution in Polymers

Sychrovský V.[1], Šebera J.[1], Fukal J.[1]

[1] *Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Flemingovo náměstí 2, 166 10, Praha 6, Czech Republic*

Novel unorthodox approaches employing deep learning techniques can nowadays predict protein's 3D structure owing to extensive training based on plausible x-ray structural data. However, native state of a biochemically active polymer is rather structure-dynamical. A validation of 3D polymer's motives beyond static snapshots due to x-ray is obviously desirable. Their proper description within a training data set might enhance applicability of novel structure-predicting tools. Molecular dynamic simulation can extend structural picture of biopolymers towards structure-dynamic one; however, its accuracy is often questioned due to deficiencies of available force fields. Hence, only molecular dynamics validated against adequate *in liquid* experiment can illuminate true state of a biopolymer. We will introduce newly developed theoretical method for structure-dynamic interpretation of NMR spectra in polymers where a model of static 3D molecular structure is replaced by probability distribution for NMR-assigned geometrical parameter(s). So far omitted dynamical component in a training data could be resolved in this way.

L1-04

Advanced Computational Protocol for Atomistic Understanding and Modulation of Insulin Binding to Insulin Receptor

Yurenko Y. P.[1], Muždalo A.[1], Černeková M.[1], Řezáč J.[1], Fanfrlík J.[1], Žáková L.[1], Jiráček J.[1], Hobza P.[1], Lepšík M.[1]

[1] *Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague, Czechia*

The binding of insulin to its receptor (insulin receptor, IR) is mediated by an extensive network of non-covalent interactions. Their quantitative characterization is difficult due to the huge size of the interface, its flexibility and thus also imprecise structural information. Major breakthroughs in the cryo-EM methodology yielded several structures of insulin bound to its receptor (IR) at sufficient resolution (3.2 Å) for atomistic modeling and structure-based drug design. Taking this into account, we have post-processed the recent cryo-EM structure of insulin/IR complex with a hierarchical computational protocol developed in this work. It entails molecular dynamics (MD), fragmentation and quantum chemical (QM) and faster molecular mechanics (MM) calculations to identify interaction "hotspots" in the IR primary site. The identified interaction hotspots in insulin Sites 1a and 1b (Ile A2, Glu A4, Tyr A19, Cys B7, Val B12, Glu B13, Tyr B16, Phe B24 and Phe B25) are in excellent agreement with the available experimental data. The energetic description of individual non-covalent interactions at the interface by the SQM and MM methods was checked on smaller fragments (amino acid dimers and trimers, up to 200 atoms) against DFT-D3/COSMO calculations. Both methods satisfactorily reproduce weaker contacts and strong NH…O hydrogen bonds but give a different ordering of individual residues. This proves a suboptimal description of the interactions using the MM method as compared to SQM. The SQM-based computational protocol developed in this work was validated against experimental and higher-level QM data and found applicable to very large and extremely flexible protein-protein interfaces.

SESSION 2

Tools

L2-01

SeqUIa: a software platform for GUI based next-generation sequencing data analysis

Bystry V.[1], Juraskova K.[1], Demko M.[1], Jugas R.[1], Hejret V.[1], Trachtova K.[1,2], Blavet N.[1], Pokorna P.[1], Palova H.[1], Alexiou P.[1]

[1] Central European Institute of Technology, Masaryk University, 60177 Brno, Czech Republic.
[2] Christian Doppler Laboratory for Applied Metabolomics (CDL-AM), Medical University of Vienna, 1090 Vienna, Austria.

Next-generation sequencing (NGS) is the molecular diagnostic technology of the future, which with its significantly dropping costs, started recently to replace the standard diagnostic laboratory techniques. The well-known bottleneck of NGS-based diagnostics is the subsequent bioinformatics analysis, which usually consists of multiple complex steps requiring specific algorithms and settings. Many existing software tools cover only some parts of NGS diagnostics or specialize in specific types of NGS diagnostics, but to the best of our knowledge, there is no generic tool for setting and processing the complete bioinformatics analysis of NGS diagnostics.
Here we present SeqUIa, a software platform to organize, control, and run the NGS bioinformatics analysis through an intuitive web-based GUI. The software handles the whole process, from sample registration all the way to the report visualization. SeqUIa is a full-fledged laboratory information system for NGS-based molecular diagnostics, which includes the means to run the analytical workflows effortlessly. SeqUIa works with Snakemake workflows to ensure reproducible and scalable data analyses. Standard NGS analyses are included in the system, but SeqUIa allows for an easy “plugin” of any additional Snakemake workflow. The software features a 2-tier architecture designed for secure work with sensitive data and flexible deployment to various computational resources.

L2-02

Deep Learning the binding patterns of RNA-binding proteins using ENNGene

Chalupová E.[1,2], Vaculík O.[1,2], Poláček J.[3], Jozefov F.[3], Majtner T.[2], Alexiou P.[2]

[1] *Faculty of Science, National Centre for Biomolecular Research, Masaryk University, Brno, Czechia*
[2] *Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czechia*
[3] *Faculty of Informatics, Masaryk University, Brno, Czechia*

Human genome encodes more than 1500 known RNA-binding proteins (RBPs). Research in the last decades clearly showed that the protein-RNA binding is driven by a combination of multiple factors. Although the RNA sequence is often one of the major binding determinants, other aspects such as site availability, binding competitors, or multimerization of the protein are also crucial. Whether the binding event will occur or not thus can not be easily predicted by classical motif search.
Deep Learning (DL) models now commonly outperform the older Machine Learning (ML) and other methods in RBP target site classification and present a prospective approach able to learn complex patterns given enough data. We show that we can easily reach or even outperform the state-of-the-art results using the ENNGene tool by quickly searching for optimal network architecture and including evolutionary conservation as an additional input feature.
ENNGene is a tool developed in our lab that allows training of custom Convolutional or hybrid Convolutional-Recurrent Neural Networks on any Genomic data through a Graphical User Interface. The tool allows multiple streams of input information, including sequence, evolutionary conservation, and predicted RNA secondary structure. ENNGene deals with all steps of data preprocessing, model training, and evaluation, exporting useful metrics and graphs. To facilitate interpretation of the predicted results, Integrated Gradients provide the user with a graphical representation of an attribution level of each nucleotide of a predicted sequence.

L2-03

HiC-TE: a pipeline for HiC data analysis in the context of repeats and genome organization

Lexa M.[1,2], Cechova M.[1], Son Hoang Nguyen[1], Jedlicka P.[2], Kubat Z.[2], Hobza R.[2], Kejnovsky E.[2]

[1] *Faculty of Informatics, Masaryk University, Brno, Czech Rep*
[2] *Institute of Biophysics, Czech Academy of Sciences, Brno, Czech Rep*

High-throughput chromosome conformation capture (Hi-C) has become a well established sequencing-based method to detect physical proximity of DNA segments in nuclei, with thousands of Hi-C experiments now available in public repositories (e.g. NCBI Sequence Read Archive (SRA)). We combined this data with increasingly precise public plant reference sequences and tools that characterize the repetitive fraction of genomes, such as Tandem Repeat Finder, PlantSat database, TE-greedy-nester or Repeat Explorer 2. We built a Nextflow pipeline with two main inputs, a SRA sequencing run ID and a corresponding reference genome. The pipeline processes the Hi-C reads, maps and clusters them to identify Hi-C pairs that can be attributed to specific repeat classes, quantifying them within and between repeat families. Results of the analysis are conveniently visualized as heatmaps, circular chromosome plots and exported as data tables. The pipeline is available for use via a gitlab repository (https://gitlab.fi.muni.cz/lexa/hic-te). First experiments show biologically important interactions of ribosomal DNA clusters or centromeric repeats that are clearly visible in most plant species. We discovered that LTR retrotransposon families with high interaction rates are often species-specific. This pipeline represents a novel and reproducible way to analyze the role of repetitive elements in the 3D organization of genomes.

L2-04

Unsupervised automated population detection and immunophenotypisation tool for analysis of multiparameter flow cytometric data

Podolská T.[1], Sieger T.[2], Fišer K.[1,3]

[1] CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic.
[2] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[3] Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic

The multiparameter flow cytometry (MFC) is an important part of Acute lymphoblastic leukaemia diagnostics and disease monitoring. The MFC data are conventionally analysed manually by experts in the field. However, such analysis is subjective and relies on high level of expertise. While a number of computational tools for MFC have been developed, no fully automated unsupervised method spanning population identification, phenotypisation and reporting has been proposed yet.
Here we propose a computer-assisted analytical protocol for analysis of MFC files. The analysis consists of four steps: 1) preprocessing, 2) clustering by MFC tailored hierarchical clustering analysis (Fišer et al., 2012), 3) population selection, 4) phenotypisation of the detected populations (Sieger et al., unpublished work). All four steps are performed without human input by issuing a single command.
The population identification step is based on hierarchical clustering and subsequent cluster selection. We developed a cluster selection approach based on cluster compactness evaluation metric (Sieger et al., unpublished work). In our approach the compactness of each dendrogram node is evaluated. In each step a cluster with the highest compactness value is selected and all its parent and children clusters are excluded from further selection. This selection is repeated until all cells belong to a cluster. Thus, the clusters are identified at different dendrogram tree heights.
We tested the proposed automated analytical protocol on a set of fcs files (n = 101) from the same 8-parameter antibody panel. This was a cohort of patients indicated as suspectly having childhood acute leukemia. The resulting immunophenotypes of leukemic populations represented by quantitative metrics were selected and compared with phenotypes reported by expert cytometrists. For six out of eight markers measured in the analysed files we achieved a statistically significant agreement ($P < 0.01$) of phenotypes reported by our automated protocol and phenotypes reported by expert, according to Spearman's correlation coefficient.

Our protocol for an automatic MFC data analysis can produce biologically relevant results corresponding with expert reports. Our approach is unsupervised, automatic, fast and therefore is better scalable.

This work was supported by GAUK number 352922.

L2-05

NarCoS: Integrating genomic surveillance of SARS-CoV-2 positive clinical samples in Slovak republic

Szemes T.[1,2,3,4], Sládeček T.[1], Sitarčík J.[1,3,5], Krampl W.[1,4], Hekel R.[1,3,4], Gažiová M.[5], Böhmer M.[1,2], Kaliňáková A.[2], Staroňová E.[2], Rusňáková D.[1,2,4], Sedláčková T.[1], Budiš J.[1,3]

[1] *Comenius University Science Park, Bratislava, Slovakia*
[2] *Public Health Authority of the Slovak Republic, Bratislava, Slovakia*
[3] *Slovak Center of Scientific and Technical Information, Bratislava, Slovakia*
[4] *Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia*
[5] *Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*

More than two years have passed since the World Health Organisation announced the global pandemic of SARS-CoV-2 virus with major impact on healthcare, economy in virtually all countries around the world. Genomic variability of the coronavirus led to emergence of novel phylogenetic variants with increased infection rate and partia evasion from specific immunity. Therefore, one of key tasks in pandemic management has been the genomic surveillance of clinical COVID-19 cases in all countries. Although our group had sequenced the first six clinical case as early as March 2020, systematically organised weekly sequencing has not started until March 2021. The sequencing efforts in Slovakia are coordinated by the national Public health authority. Based on lower number of cases in early 2021, the original planned sample throughput was project up to 500 samples per week and four laboratories were involved. With the emergence of Delta and Omicron variants the required sample number has grown almost four times. This led to increase in analysed samples in participating laboratories and adding of further two sequencing laboratories, but with limited experience with NGS technology and absent experience with bioinformatics processing. To ensure reliable per case analysis of clinical samples in six Illumina systems based sequencing laboratories, metadata transfer automation, unified variant calling and interpretation and batch upload to GISAID and ENA database as well as fast reporting to TESSY database, we developed an web based integrated information system for sequencing management, data and metadata transfer and automated batch reporting and uploading to relevant databases. Uniquely tuned in-house variant calling pipeline allowed us to unify the analysis of all samples from Slovakia. The name of the system is NarCoS. Aggregate results visualisations are used for reporting purposes for Healthcare ministry pandemic commission for prompt situation assessment and management. We continue to further develop the system, currently focusing on the integration of sequencing based wastewater monitoring for SARS-CoV-2 variants.

SESSION 3

Small molecules

L3-01

Towards the interpretation of tandem mass spectra with self-supervised machine learning

Bushuiev R.[1,2], Pluskal T.[1]

[1] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo nám. 2, 160 00 Praha 6, Czech Republic*
[2] *Czech Technical University in Prague, Faculty of Information Technology, Thákurova 9, 160 00 Praha 6, Czech Republic*

The identification of unknown molecules is the main bottleneck in biological and biochemical applications of mass spectrometry. State-of-the-art computational approaches deduce desired structural properties of molecules from their tandem mass spectra, relying on previously annotated spectral libraries. However, the diversity of mass spectrometry conditions and limited scope of the spectral libraries limit the capabilities of such methods. In our project, we propose to break the limitation with the paradigm of self-supervised machine learning. Firstly, we extract a high-quality subsample of unannotated tandem mass spectra from the MassIVE repository (hundreds of millions of spectra). Secondly, we use the obtained data to pre-train MSBERT - a version of the prominent BERT language model adapted to operate on mass spectra. For this purpose, we employ an analog of the "masked language model" training objective, where we hide random peaks in the spectra and train the model to predict the hidden peaks. Thirdly, we fine-tune MSBERT end-to-end to predict structural properties of small molecules (e.g. number of Carbon atoms or presence of Nitrogen), using annotated spectral libraries as training data. In preliminary experiments, we have shown that self-supervised pre-training helps MSBERT to solve the downstream tasks. It implies that, similarly to BERT in natural language, MSBERT has the potential to become a working horse for a wide range of mass spectrometry problems. Therefore, we are currently designing an advanced architecture of the model and estimating its optimal training hyperparameters to shift its inductive bias from the natural language to mass spectrometry.

L3-02

Machine learning estimated docking scores

Clarová K.[1]

[1] *VŠCHT Praha*

Docking programs, used for virtual screening of molecular docking, are rather difficult to operate and time and hardware demanding. In our project, we sought to develop methods for predicting docking scores using machine learning (ML) and deep learning (DL) and thus address mentioned problems and the need to use specialized docking software. The necessary data were obtained using the MOE docking software; the dataset was used for designing, implementation and testing machine learning methods. We compared the regression models of ML methods and new approach in natural language processing - Transformers architecture trained on a body of chemical data presented as SMILES strings.

L3-03

Feature interrelation profiling

Čmelo I.[1], Voršilák M.[1], Dehaen W.[1], Svozil D.[1]

[1] *UCT Prague, Technická 5, 166 28 Prague 6 – Dejvice*

Chemical structures are routinely represented by vectors of their binary features, such as structural fingerprints. The feature vectors often serve as a basis for direct structure comparisons in similarity searches and visualizations, or as an input to more advanced methods and models. Feature interrelation profiling (FIP) is a generic information-theory based extension of these feature vector core concepts. FIP uses various forms of Kullback–Leibler divergence matrices to store information about feature co-occurrence probabilities within various sets of chemical structures, thus forming their interrelation profiles.

The resulting interrelation profiles can be used to directly compare chemical structures and their sets based not only on the amount of individual shared features, but also on how consistent is a given feature combination with a reference interrelation profile. To quantify said consistency of given feature combination against a reference interrelation profile, we propose a measurement of "relative feature tightness" (RFT). For a given chemical structure or their set, RFT weighs all observed feature co-occurrences against their co-occurrence scores within a given reference interrelation profile. RFT yields a scalar value that describes the overall match in present feature combinations, much like Tanimoto similarity describes the overall match in present features.

This method of quantifying the relative "favorableness" of feature combinations, as well as the interrelation profiles themselves, present additional layers of information that can be easily used in similarity searches, visualizations and other methods as a complement to the standard feature vectors. This poster shows the theoretical background of FIP, its uses for visualization and its pilot application on assessing synthetic accessibility of chemical structures, with performance close to that of dedicated methods like SAScore and SYBA. [1]

References

[1] I. Čmelo, M. Voršilák, D. Svozil, J Cheminform 13, 3 (2021).

L3-04

Circular fingerprint inversion: an algorithmic approach

Dehaen W.[1], Čmelo I.[1], Karlova A.[2], Svozil D.[1]

[1] *Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Technicka 5, 166 28, Prague, Czech Republic*
[2] *Department of Computer Science University College London London, United Kingdom*

Molecular structures can be conveniently and concisely converted to molecular fingerprints, which allow fast database searching, model building and other tasks. However, they are not a full molecular representations, and typically there is some degree of information loss when converting a molecular structure to its corresponding fingerprint. In other words, converting a fingerprint back to its parent structure is a non-trivial and often impossible task if a structure fingerprint dictionary lookup is not possible.

One of the most commonly used families of fingerprints are circular fingerprints, such as the Morgan fingerprint, also referred to as the Extended-Connectivity Fingerprints(ECFP). These fingerprints encode chemical environments within a defined topological distance from each atom in the structure. Conversion to a circular fingerprint was considered to be a one-way operation, and even a way to „encrypt“ proprietary structures, but recently Le *et al* used a neural network based method to show a significant percentage (10-50%) of ECFP can be inverted. This surprising and powerful result does have some limitations: the neural architecture used is quite involved and requires a non-trivial amount of training resources, including a training set of 1.4 million drug-like molecules. Additionally, a different model needs to be trained for different fingerprint settings, such as radius and bitvector length.

To address these limitations, we investigated an algorithmic approach to fingerprint reconstruction. Briefly, the unknown molecule is built atom per atom in a depth-first type search, at every stage matching the intermediate molecule to the fingerprint and backtracking when there is an on-bit in the fingerprint that does not match the target fingerprint. This strategy managed to recover a large percentage of structures and had the advantage of not requiring any training data, meaning biases in a training set (such as only drug-like molecules, only synthesizable molecules etc) are not a factor. No training time is required, so inversion at given settings can immediately be started. To test this idea, we reconstructed molecules from the GDB, a database that encodes all possible structures of a given heavy atom count satisfying certain minimal criteria (only heavy atoms C,N,O,F, valence rules, stability rules etc). Our prototype approach reconstructed 100% of gdb-7, 99.3% of gdb-8, 95.1% gdb-9, 88.5% of gdb-10 and 82.4% of gdb-11 from their ECFP6(2048).

L3-05

On the Importance of Physically Correct Models for Describing Protein/Ion/Ligand Binding

Lepšík M.[1*]

[1] *Institute of Organic Chemistry and Biochemistry (IOCB) of the Czech Academy of Sciences, Prague, Czech Republic*
[*] *lepsik@uochb.cas.cz*

Understanding protein-ligand binding in atomistic details is the key to success in structure-based drug design. Herein, I review i) the progress of corrected semiempirical quantum mechanics (QM)-based scoring function in sampling, ranking and virtual screening1,2 and ii) application of effective electronic polarization scheme for classical molecular dynamics (MD) which helps explain a rare oligosaccharide conformer in lectin/calcium/carbohydrate complex.3 In summary, developing and applying physically correct models of protein-ligand binding heads toward an unrivaled qualitative enhancement of the predictive power of computer-aided drug design.

References

[1] Lepšík, M. et al., The Semiempirical Quantum Mechanical Scoring Function for In Silico Drug Design. *ChemPlusChem* **2013**, *78*, 921 – 931.
[2] Pecina, A. et al., SQM/COSMO Scoring Function: Reliable Quantum-Mechanical Tool for Sampling and Ranking in Structure-Based Drug Design. *ChemPlusChem* **2020**, *85*, 2362 –2371.
[3] Lepsik, M. et al. Induction of rare conformation of oligosaccharide by binding to calcium-dependent bacterial lectin: X-ray crystallography and modelling study. *Eur J Med Chem*. **2019**, *177*, 212-220.

L3-06

Multiple instance learning: a new method for 3D QSAR modelling

Matveieva M.[1,2], Zankov D.[3,4], Nikonenko A.[2], Madzhidov T.[3], Polishchuk P.[2]

[1] *Department of Informatics and Chemistry, University of Chemistry and Technology (VŠCHT) Prague, Czech Republic, matveiem@vscht.cz*
[2] *Institute of Molecular and Translational Medicine, Faculty of Medicine and Dentistry, Palacky University Olomouc, Olomouc, Czech Republic*
[3] *Laboratory of Chemoinformatics and Molecular Modeling, A. M. Butlerov Institute of Chemistry, Kazan Federal University, Kremlyovskaya 29, 420111 Kazan, Russia.*
[4] *Laboratory of Chemoinformatics, Institute Le Bel, University of Strasbourg, B. Pascal 4, 67081 Strasbourg, France*

QSAR approaches based on the use of 2D descriptors have wide practical applications in drug discovery for designing potentially bioactive molecules; however, they do not account for important information contained in 3D structures of molecules. The major limitation in constructing models using 3D descriptors is the choice of a putative bioactive conformation, which affects the predictive performance. To overcome this problem multi-instance (MI) learning approach considering multiple conformations in model training could be applied. In this study, we implemented several multi-instance algorithms, both conventional and based on deep learning, and investigated their performance. We compared the performance of MI-QSAR models with those based on the classical single-instance QSAR (SI-QSAR) approach in which each molecule is encoded by either 2D descriptors computed for the corresponding molecular graph or 3D descriptors issued for a single lowest energy conformation. The calculations were carried out on 175 data sets extracted from the ChEMBL23 database. It is demonstrated that (i) MI-QSAR outperforms SI-QSAR in numerous cases and (ii) MI algorithms can automatically identify plausible bioactive conformations.

L3-07

Nature-Inspired Antivirals with Distinctive Mechanisms of Action: Focus on HIV and SARS-CoV-2

Ntie-Kang F.[1], Simoben C. V.[1], Lobe M. M. M.[1,2], Eni D. B.[1], Babiaka S. B.[1], Duran-Frigola M.[3], Tietjen I.[4], Efange S. M. N.[1]

[1] Department of Chemistry, University of Buea, Cameroon
[2] Institute of Medical Research and Medicinal Plants, Ministry of Scientific Research and Innovations, Yaounde, Cameroon
[3] Ersilia Open Source Initiative, Cambridge, UK
[4] Small Molecule Discovery and Pharmacognosy Group, Vaccine and Immunotherapy Center, The Wistar Institute, Philadelphia, PA 19104, USA

This project is focused on nature-based discovery of antiviral agents that target putative drug targets in the human immunodeficiency virus (HIV) and the severe acute respiratory syndrome (SARS) coronavirus disease 2019 (COVID-19) virus (SARS-CoV-2) for which screening procedures have lately been established. The presenter has previously developed the web-accessible African Natural Products Database (ANPDB). The application aims to set up a cloud-based computing platform with an *in silico* pipeline coupling artificial intelligence/machine learning (AI/ML) with physics-based methods (e.g. molecular docking, molecular dynamics, etc.) and standard ligand-based statistical search methods (e.g. quantitative structure-activity relationships (QSAR), sub-structure and similarity searches along with pharmacophore-based methods). The *in silico* (virtual) hits from the pipeline, starting from the entire natural products library of all medicinal plants growing in Africa, will be screened in a panoply of assays with the aim of identifying compounds inhibiting vital targets like the SARS-CoV-2 spike protein and proteases, as well as vital ion channels and latency reversal in HIV. Lead expansion aimed at exploring the chemical space around the identified compounds from the screens is also planned, along with the screening of the entire in-house library of synthetic mimics of natural products from the applicant's laboratory. In addition to targeting these vital viral proteins with synthetic analogues aiming at nanomolar range inhibitors that could be taken to *in vivo* experiments, the project aims at transferring the know-how in assay development and cloud computing to our African laboratory.

L3-08

Techniques for improving optimization performance of molecular generators

Pešina F.[1,2], Svozil D.[1,2]

[1] *CZ Openscreen*
[2] *UCT Prague*

Over the last few years, deep neural networks became an indispensable tool in de-novo computational drug design via molecular generators. A wide range of architectures and optimization algorithms have already been tested with promising results. In addition to that, there are other features that can potentially improve generator performance. Two of those promising features are Transfer learning and Initial population. Both were already used in various setups, however a multi-generator comparison aimed to evaluate the contribution of these features to optimization enhancement under standardized conditions is still missing. In this work, we performed such a comparison using several deep neural networks-based generators and a set of standardized benchmarks from GuacaMol benchmarking platform. We report that Transfer learning significantly enhanced the optimization results across all tested generators and majority of benchmarks both with regard to achieved results and also the amount of time necessary to achieve those results. Use of Initial population also led to optimization improvement in some cases, however to a much lesser extent.

SESSION 4

Sequences

L4-01

Prediction of terpene synthase activity using self-supervised deep learning

Samuchevich R.[1,2,3], Čalounová T.[1,4], Bushuiev R.[1,5], Tajovská A.[1], Šivic J.[2], Pluskal T.[1]

[1] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences*
[2] *Czech Institute of Informatics, Robotics and Cybernetics (CIIRC CTU)*
[3] *Faculty of Chemical Technology, University of Chemistry and Technology*
[4] *Faculty of Science, Charles University in Prague*
[5] *Faculty of Information Technology, Czech Technical University in Prague*

Terpene synthases (TPSs) are enzymes responsible for the biosynthesis of the largest class of natural products, including widely used flavors, fragrances, and first-line medicines. The amount of available TPS protein sequences is increasing exponentially due to rapid advances in high-throughput sequencing. However, characterizing the function of each TPS requires challenging and time-consuming experiments as well as significant domain expertise. To help overcome these challenges, we employ self-supervised deep learning models to understand the sequence space of TPSs and make predictions regarding the substrates and reactions catalyzed by yet-uncharacterized TPS enzymes. Our model can now outperform established bioinformatic methods based on hidden Markov models for the prediction of TPS substrates. To facilitate the development of novel predictive methods for TPS characterization, we also assembled a curated database of ~1,500 characterized TPS-catalyzed reactions, which is currently the largest such dataset available.

L4-02

Using ChIP-nexus to decipher the architecture of transcription factor complexes

Převorovský M.[1], Marešová A.[1], Hradilová M.[2]

[1] Laboratory of Microbial Genomics, Department of Cell Biology, Faculty of Science, Charles University, Prague, Czechia
[2] Genomics and Bioinformatics Core Facility, Institute of Molecular Genetics of the ASCR, v.v.i., Prague, Czechia

Transcription factors are key regulators of gene expression. They often form complexes with transcriptional coactivators and/or corepressors. Moreover, individual genes are often regulated by multiple such complexes. Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a popular method for mapping transcription factor binding sites in a genome-wide manner. However, ChIP-seq typically lacks the resolution required for accurate determination of the exact positioning of regulatory proteins on DNA. This, in turn, limits our ability to determine the relationships between different regulators acting on common target genes. Therefore, methods such as ChIP-exo or ChIP-nexus have been developed to increase the spacial resolution of the genome-wide ChIP assay by employing DNA digestion of regions not protected by DNA-bound proteins. We have used ChIP-nexus to describe the transcription factor network regulating lipid metabolism genes in the fission yeast *Schizosaccharomyces pombe*. We will present a complete workflow for ChIP-nexus data acquisition, processing and analysis, together with the key biological conclusions.

L4-03

Transformer Language Models for Genomic Sequences

Šimeček P.[1], Martínek V.[1], Čechák D.[1]

[1] *Central European Institute of Technology (CEITEC) MU, Brno*

Since long short-term memory (LSTM) architecture, neural networks have been proven helpful for natural language processing (NLP) tasks. But what about the language of genomic sequences written in a four-letter alphabet? In the last few years, researchers have demonstrated that neural networks can be used to learn to identify functional elements in genomic sequences, and that the resulting models can predict the function of previously uncharacterized genomic regions.

The "revolution" had come in 2018 with the transformer architercure having the ability to detect complex dependencies between elements of a series thanks to a mechanism of attention or self-attention. This talk will review known pre-trained models, explain how they can be fine-tuned to a specific task, and show the usage on a couple of genomic benchmarks. We will demonstrate that the embeddings (=numeric representations) of genomic sequences contain additional information that the model has not been implicitly taught. We conclude by discussing practical aspects like the hyperparameter search and deployment to Hugging Face Models / Spaces.

L4-04

Modelling of Ago2 Binding using CLASH

Hejret V., Varadarajan N. M., Klimentova E., Giassa I., Vanacova S., Alexiou P.

Argonaute proteins play a central role in the regulation of RNA stability and translation, via a targeting process mediated by small RNA ‘driver’ sequences that drive the protein to its targets. Ago2, the major mammalian Argonaute protein, is known to primarily associate with microRNAs, a family of small RNA ‘driver’ sequences, and identify its targets via a ‘seed’ mediated partial complementarity process. Despite a number of experimental and computational studies that have approached the question of Ago2 targeting, a clear experimental dataset of Ago2 ‘driver’ - target interactions has not been available to date. We present the first Ago2 CLASH experiment, which produces thousands of Ago2 target sites supported by chimeric reads that include together fragments of the ‘driver’ and the target sequence. Using a novel analysis pipeline we report thousands of Ago2 target sites driven by microRNAs, but also a substantial number of Ago2 ‘drivers’ derived from fragments of other small RNAs such as tRNAs, snoRNAs, rRNAs and others. We have produced machine learning based computational models that efficiently predict the binding potential for each of these ‘driver’ classes, and experimentally validate a number of interactions. We expand our knowledge of its ‘driver’ repertoire and potential function in development and disease. This research was funded by Grantová Agentura České Republiky, 19-10976Y Grant to P.A.

L4-05

Prediction of sequence divergence from the quality of mapping

Kovacova V.[1], Lässig M.[1]

[1] *IBP, University of Cologne, Zülpicher Str. 77, Cologne, Germany*

Read mapping, one of the first steps in sequencing data processing, can significantly influence downstream analysis results. Simulated sequencing libraries with different levels of sequence divergence mapped to the same reference genome display drop-off of mapped reads from tenths to tens of per cent. For example, if the sequence divergence is close to seven per cent, we see a considerable decrease of mapped reads under the default setting of the STAR aligner.

Correctly adjusting a reference genome and testing the mapping tool's parameters to decrease false findings is crucial, mainly in microbial studies. On the other hand, known proportions of unmapped reads under defined mapping parameters can inform us about sequencing divergence and evolution of given (sub)species.

To compare the results of simulated data sets with natural sequences, we used *B. subtilis*'s subspecies *spizizenii* and s*ubtilis, where speciation led to ne*arly 7% sequence divergence.

L4-06

Interpreting uncertainty in differential expression with DESeq2

Modrák M.[1]

[1] *Bioinformatics Core Facility, Institute of Microbiology of the Czech Academy of Sciences*

While differential expression analysis is often reported primarily as a list of genes where the hypothesis of no/low change can be rejected, further insights may be gained by examining the reported uncertainty of the log fold change. In a simulation study we investigate the coverage properties of confidence intervals derived from the reported standard errors in DESeq2. We note that with the most straightforward normal approximation, the tail (> 90%) confidence intervals are overly narrow and have coverage around 1 – 5 percentage points lower than nominal across a range of settings. Using a student-t approximation can overcorrect and result in too wide confidence intervals. Moreover, if only differentially expressed genes are selected for reporting, an adjustment of the confidence intervals for multiple comparisons is necessary to avoid overconfident claims. We note that this mild discrepancy is completely compatible with p-values from DESeq2 providing tight control of false positives. The same simulation setup also allows us to judge how closely the empirical Bayes shrinkage methods implemented in DESeq2 approximate an exact Bayesian posterior. We note that while the approximation is imperfect, Bayesian interpretation of uncertainty intervals from DESeq2 as “likely” containing a “true” value of the fold change is not completely unwarranted, especially if higher number of replicates and/or a student-t approximation is used.

L4-07

GC content of transposons and of their (animal) host genomes

Symonova R.[1,2], Kubečka J.[2], Matoulek D.[3]

[1] *Technical University of Munich*
[2] *Institute of Hydrobiology, Biology Centre CAS, České Budějovice*
[3] *University of Hradec Kralove, Czech Republic*

Fish genomes are AT/GC homogenous in comparison with heterogeneous genomes of mammals and birds. The heterogeneity in mammals is apparent on DNA sequences as well as on chromosomes upon a suitable staining. With our finding that both extant genera of ancient ray-finned fish gars possess AT/GC heterogenous genomes similarly as mammals do, we initiated a search for possible reasons of this compositional heterogeneity. In this effort, we developed a Python tool segmenting genomes assembled to the chromosome level according to their GC% and repetitive content (soft-masked DNA) and plotting these values in a single plot. The plot is represented by a colored profile of GC% values along the chromosomes, where the color represents the gradient between the fully repetitive DNA (green) and fully non-repetitive DNA (red) with the sliding window of the size 1 kb as the default but freely adjustable. Our survey across numerous freshwater fish assemblies with this tool showed the repetitive fraction as the potentially AT/GC homogenizing factor in lower vertebrates. At this stage, the quality of the soft-masking procedure is the most crucial step. Otherwise, this tool can be used in any other group with genomes assembled to the chromosome level, e.g. plants and fungi, to explore the role of the repetitive fraction in other major lineages beside vertebrates.

The study was supported by the European Union within ESIF in frame of Operational Programe Research, Development and Education (project no. CZ.02.1.01/0.0/0.0/16_025/0007417 administrated by the Ministry of Education, Youth and Sports of the Czech Republic)

Poster session

Monday, 13. June

Poster session is sponsored by the
National Infrastructure of Chemical Biology.

CZ OPENSCREEN
National Infrastructure for Chemical Biology

P-01

The use of a targeted RNA sequencing-based approach for the detection of clinically relevant fusion genes in pediatric cancer patients

Al Tukmachi D.[1], Veselá P.[1], Trachtová K.[1], Bystrý V.[1], Tichý B.[1], Slabý O.[1,2,3]

[1] *Central European Institute of Technology (CEITEC), MU, Brno*
[2] *Department of Biology, Faculty of Medicine MU, Brno*
[3] *Comprehensive Cancer Care Department, Masaryk Memorial Cancer Institute and FM MU, Brno*

Childhood cancers are among the rare diseases with an annual incidence of around 350-400 new cases in the Czech Republic and represent the second most common cause of death in this age group. With significant advances in molecular profiling techniques and their successful implementation within various precision oncology programs and pediatric pan-cancer profiling initiatives, different NGS-based methods are becoming more routinely used in a clinical setting. Fusion gene analysis is of great importance for diagnostic and prognostic stratification and therapeutic planning among these methods. Between September 2019 and May 2022, 243 pediatric cancer patients underwent fusion gene analysis using targeted RNA sequencing. In 210 cases, the analysis was carried out as a part of a complex tumor profiling within a precision oncology program. Sequencing libraries from both fresh-frozen and FFPE tissue were prepared using TruSight RNA Pan-Cancer Panel (Illumina) covering 1385 cancer-associated genes and sequenced on NextSeq 500 platform using NextSeq Mid Output Kit (150 cycles) (Illumina). Raw reads were quality checked with the FastQC package (version 0.11.9). Adapter sequences were identified and trimmed with the Trimmomatic tool. Trimmed reads were then mapped to reference genome hg38 using a STAR aligner with parameters set to allow fusion gene detection. Mapping quality was checked using QualiMap and Picard tools. Fusion genes were identified using Arriba and STARfusion. Visual verification of identified fusion genes was performed using IGV software. In 26 % of cases, a clinically relevant fusion gene was identified. The most commonly found fusions included known drivers of sarcoma tumorigenesis, such as *EWSR1-FLI1*, *PAX3-FOXO1*, and *SS18-SSX1/2*. The second largest group consisted of fusion genes associated with CNS tumors, mainly low-grade gliomas, e.g., *KIAA1549-BRAF* or other MAPK-activating fusions. Approximately one-third of identified fusion genes were therapeutically actionable. In one case, a novel fusion gene *DVL3-TFE3* was identified, relevant to the renal cell carcinoma diagnosis of the analyzed patient. Targeted RNA sequencing has proven to be a sensitive and feasible strategy contributing to patient stratification and treatment selection. Its ability to detect both clinically established and novel fusion

genes is superior to PCR-based approaches that show only a limited throughput. Further use of this method will allow a better understanding of cancer biology.

P-03

PENGUINN-RNA: prediction of RNA G-quadruplexes using interpretable Neural Networks

Bhagat K.[1], Giassa I.[1], Alexiou P.[1]

[1] *Central European Institute of Technology, Brno, Czech Republic*

G-quadruplexes (G4s) are non-canonical structures of nucleic acids that have gained increasing interest due to their involvement in a series of biological processes. While the first DNA G4 was identified more than 30 years ago, RNA G4s became known two decades ago. Since then there is accumulating evidence for their importance in cellular mechanisms, including translation regulation, telomere maintenance, and alternative splicing. Here we present PENGUINN-RNA, a machine learning method able to predict RNA G4s based on raw RNA sequence and highlight the regions of the sequence that contribute to the formation of the G4 structure. The trained model is available online and is also accessible through a user-friendly interface that can calculate the G4-forming propensity of user-submitted RNA sequences.

P-05

Computational method for the detection of microsatellite instability in tumor tissue samples

Budiš J.[1,2,3], Sýkora M.[4], Kucharík M.[1,2], Krampl W.[1,2], Pös O.[1,2], Styk J.[1,2,5], Radvánszky J.[2,6,7], Szemes T.[1,2,6]

[1] Geneton s.r.o., Bratislava 84104, Slovakia
[2] Science Park, Comenius University in Bratislava, 84104, Slovakia
[3] Slovak Center of Scientific and Technical Information, Bratislava 81104, Slovakia
[4] Faculty of Informatics and Information Technologies, Slovak university of technology in Bratislava, 84216, Slovakia
[5] Faculty of Medicine, Comenius University in Bratislava, 81372, Slovakia
[6] Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava 84215, Slovakia
[7] Institute of Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Bratislava 84505, Slovakia

Recent advances in sequencing technologies have enabled affordable testing for various genetic diseases. Their application in the field of oncology has great potential for early detection and monitoring of tumor progression. A promising way is the thorough characterization of microsatellites, where the detection of their instable forms can be used as a cancer biomarker.
We compared sequenced DNA fragments isolated from tumour and control tissues. We first mapped them to the reference human genome. Then, thousands of selected microsatellite loci across the genome were thoroughly characterised to detect typical anomalies of unstable forms. The changes were aggregated across the analysed loci to obtain compact features for each sample. Finally, the features were analysed with classification methods to distinguish between tumor and control tissue.
We show that microsatellite instability is readable in our tumour samples. The method has therefore the potential to automate the detection and characterization of ongoing oncology disease from the sequenced genomic data.

The presented work was supported by the Slovak Research and Development Agency (grant ID APVV-18-0319). The presented work was also supported by the OP Integrated Infrastructure within projects with ITMS codes 313011W988, 313011F988, and 313011V578, all co-financed by the European Regional Development Fund.

P-07

Bioinformatics tools for Non-Invasive Prenatal Testing

Dohnalová H.[1*], Němec M.[1], Nguyen Thi Ngoc B. L.[1], Zembol F.[1], Bittóová M.[1], Hrabíková M.[1], Koudová M.[2], Stejskal D.[2]

[1] *Laboratory of Molecular Genetics, GENNET, s.r.o.*
[2] *Department of Medical Genetics, GENNET, s.r.o.*
[*] *Hana.Dohnalova@gennet.cz*

Non-Invasive Prenatal Testing (NIPT) uses free fetal DNA (circulating cell-free DNA) present in the mother's blood. Using the whole-genome sequencing method with low coverage, the most common aneuploidies can be identified and the sex of the fetus can also be determined.
Our solution uses various bioinformatics tools. We combine three methods Defrag, SeqFF and ComboFF to determine the fetal fraction. Using WisecondorX, we calculate aberrations at the chromosomal level. Finally, we statistically evaluate and determine the sex of the fetus.
Thus, we have built a pipeline that combines various, originally stand-alone bioinformatics tools for NIPT into one easy-to-use solution for everyday clinical applications.

P-09

Paperfly: ab initio binding site reconstruction

Faltejsková K.[1,2], Vondrášek J.[1]

[1] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo náměstí 542/2, 160 00 Praha 6, Czech Republic*
[2] *Charles University, Faculty of Mathematics and Physics, Computer Science Institute, Malostranské náměstí 25, 118 00 Praha 1, Czech Republic*

The specific recognition of a DNA locus is a widely studied issue. It is generally agreed that the recognition can be influenced not only by the binding motif, but by the larger context of the binding site. In order to study the binding site including the sequential context of the binding motif, we introduce PAPerFly: the Partial Assembly-based Peak Finder, a new tool capable of reconstructing the binding site from ChIP-seq or similar experimental data.

Using a novel heuristic algorithm that utilizes approaches used in the genome assembly, Paperfly can reconstruct the unique binding sites captured in a sequencing experiment without using the reference genome. Additionally, we show that Paperfly can be combined with the standard data processing pipeline to link the unique sequences of the binding site and their respective abundance over the genome.

The source code of the tool is freely available at https://github.com/Caeph/paperfly or at https://doi.org/10.5281/zenodo.6379332.

P-11

Accessing chemical and biological datasets through SPARQL endpoints

Galgonek J.[1]

[1] *Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences, Flemingovo náměstí 542/2, 160 00 Praha 6, Czech Republic*

Many of medium to large scale biological and chemical datasets are internally stored in relational databases. This approach makes it easy to develop a dedicated web server that presents a dataset and supports data querying. Unfortunately, such a way is not usually interoperable, and it can be difficult to combine the dataset with other ones or to query multiple data sources uniformly. To address this gap, some of these datasets use the Resource Description Framework (RDF) to export their data in interoperable formats. To increase interoperability even more, some datasets support querying their data by SPARQL, the query language for RDF data. If a dataset is originally stored in a relational database, there are two basic approaches. In the first one, data are exported to an RDF form and stored in a native RDF storage supporting SPARQL querying. The disadvantage of this approach is that the data are stored twice, or migration of the used database technology is needed. The second approach is to keep the data in the relational database and use a system that allows mapping the relational data to the RDF form. This mapping is used by the system to translate incoming SPARQL queries to equivalent SQL queries that are then evaluated by the relational databases. In our work, we examine different ways how these mappings can be designed, and we compare them with each other and with the native solution. We also compare several technologies, that are typically used for these purposes in the field of biology and chemistry.

P-13

Genomic Benchmarks: A Collection of Datasets for Genomic Sequence Classification

Grešová K.[1], Šimeček P.[1], Martinek V.[1], Čechák D.[1], Alexiou P.[1]

[1] Panagiotis Alexiou Research Group, Centre for Molecular Medicine, Central European Institute of Technology, Masaryk University, Brno, Czechia.

Advances in Next Generation Sequencing have allowed quicker, cheaper and more precise sequencing of many species' genomes. However, to extract biologically significant information from obtained genomes, we need to identify various types of functional genomic elements within them. Experimental genomic annotation is expensive, time consuming and research is focused on only a few model organisms, resulting in a low number of well annotated species and a high number of sequenced, but unannotated species. Machine learning models are able to learn and generalize given information. They may be trained to learn structure of functional elements in well annotated species and then use this information to annotate new genomes from given species or even new species (i.e. cross-species annotation).
However, machine learning models are highly dependent on large amounts of high quality data to train. It is also challenging to compare the quality of different models since authors often use different datasets for evaluation and quality metrics may also be heavily influenced by data preprocessing techniques and other technical differences.
Here we propose Genomic Benchmarks: A collection of curated and easily accessible sequence classification datasets in the field of genomics. The proposed collection is based on a combination of existing datasets obtained from published papers and novel datasets constructed from mining publicly available databases. The main aim of this effort is to create a repository for shared datasets that will make machine learning for genomics more comparable and reproducible, while reducing the overhead of researchers that want to enter the field. Our repository will be especially useful for researchers with backgrounds in machine learning and statistics that aim to implement state of the art machine learning algorithms in the genomic fields but are limited by lack of domain knowledge.

P-15

Mining data from sweet cherry resequencing

Holušová K.[1], Čmejlová J.[2], Suran P.[2], Čmejla R.[2], Sedlák J.[2], Zelený L.[2], Bartoš J.[1]

[1] Institute of Experimental Botany of the Czech Academy of Sciences, Centre of the Region Haná for Biotechnological and Agricultural Research, Šlechtitelů 31, Olomouc, 779 00, Czech Republic
[2] Research and Breeding Institute of Pomology Holovousy Ltd., Holovousy 129, Holovousy, 508 01, Czech Republic

Relatively cheap sequencing produces huge amount of sequencing data. Simultaneously, the available scripts, tools and programmes allow processing gained data sets in simple ways and obtain interesting results supporting fundamental and applied research as well as breeding. Thus, genome resequencing could be valuable resource for multiple analysis. However, the data processing is labouring and time consuming, extracting all available information is not possible.
In our study, we sequenced 235 genotypes from *Prunus avium* with minimal genome coverage 20X and called SNP markers. The main goal was to associate SNPs with phenotypes scored through five years and design markers that will be used for marker-assisted selection in breeding. Besides, both the resequencing and the first basic analysis open the door for other research goals. With a basic tools we was able to detected genes responsible for particular phenotypes, find the mis-assembly in the reference genome, defined the deletion which take out MYB genes responsible for red colour of fruit, detected duplicated accessions on the basis of DNA fingerprint, designed the SSR and SNP markers for genotyping. How can this data be used more?

This work was supported by the Ministry of Agriculture of the Czech Republic (project QK1910290). Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic.

P-17

Can structural information help with phosphorylation prediction?

Kiefl Y., Gamouh G., Heinzinger M., Hoksza D., Novotny M.

[1] *Department of Cell Biology, Faculty of Science, Charles University, Czech Republic,*
[2] *Department of Software Engineering, Faculty of Mathematics and Physics, Charles University, Czech Republic,*
[3] *Department for Bioinformatics and Computational Biology, Faculty of Informatics, Technical University of Munich, Germany.*

Phosphorylation is one of the most frequent and important mechanisms the proteins have to react to changing conditions or environment. Deregulation of phosphorylation often leads to pathological conditions. Detecting phosphorylation and describing sites influenced by phosphorylation is however experimentally challenging due to the transient nature of this modification. Various sequence-based approaches have been developed to predict phosphorylation sites, but in this work we tested whether we could use the growing amount and availability of 3D structural information to predict phosphorylation sites using structural context and sequence features. We built a small set of 360 protein structures with at least one known phosphorylation site in each to train a graph neural network that will use structural and sequence features to predict phosphorylation sites. The structural information is used to give the sequence features additional context. Our initial data shows that structural features can be beneficial for prediction of phosphorylation sites, but that it has to be fine-tuned to provide competitive results. The final predictor achieves a precision of 0.47 on a class imbalance of 1/34 while retaining a low false negative rate of 0.1.

P-19

Reproducible PDX Genomic Data Analysis

Jurič B.[1], Usman M.[1], Dudová Z.[1], Peša R.[1], Stuchlík D.[1], Křenek A.[1]

[1] *Masaryk university*

Acquisition, processing, and interpretation of next-generation sequencing (NGS) data became ubiquitous in cancer research because of bringing strong evidence on biological processes related to tumor growth etc. On the other hand, besides multiple experimental methods, wide variety of computational approaches exists, yielding reproducibility and comparison of such results extremely difficult.

Our work is focused on patient-derived xenograft (PDX) mice models – in vivo human tumor implants, a well-established technique in translational cancer research and treatment selection. Besides the generic issues of all NGS data and their processing, the specific problem of PDX based data is an intrinsic mixture of human tumor and mouse tissue – those cannot be fully separated during sample preparation, the problem must be addressed computationally. Therefore, specific computational pipeline setups are also used.

The EurOPDX consortium invests significant effort into harmonization of its members data [2], including the outputs of PDX-specific NGS pipelines. However, the NGS data processing itself has not been well coordinated yet.

We present an integrated solution of several such pipelines ([1], https://github.com/jrderuiter, …) implemented in Galaxy – all the required tools were wrapped properly and connected into smoothly running workflows. The results of the workflow can be directly propagated to a testing clone of EurOPDX data portal for immediate comparison with other data already published in the Data Portal, visualization, etc.

The solution – bioinformatics pipelines and a clone of Data Portal including all visualization tools – is available as a service, where data can be uploaded and processed, or as a set of Docker containers which can be run by the user close to the data, allowing to handle non-disclosed, sensitive data etc., while still offering the full functionality.

References

[1] Woo, X. Y. *et al.* (2019). Genomic data analysis workflows for tumors from patient-derived xenografts (PDXs): challenges and guidelines. *BMC Medical Genomics*, 12.

[2] Dudová, Z., Conte, N., Mason, J. et al. The EurOPDX Data Portal: an open platform for patient-derived cancer xenograft data sharing and visualization. BMC Genomics 23, 156 (2022). https://doi.org/10.1186/s12864-022-08367-1

P-21

Decoding differentially expressed genes in artificial light at night (ALAN) induced zebrafish ovary and development of a possible major lifestyle diseases gene signature

Labala, R. K.[1,2], Khan, Z. A.[3,4], Mondal, G.[4], Vučinić K.[2], Kolář M.[2], Chattoraj A.[1*]

[1] *Biological Rhythm Laboratory, Dept. of Animal Science, Kazi Nazrul University, Asansol, 713340, India*
[2] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 14220 Prague 4, Czech Republic*
[3] *Inje University, Gimhae, South Korea*
[4] *Institute of Bioresources and Sustainable Development, DBT, Imphal, 795001, India*
[*] *asamanja.chattoraj@gmail.com*

A quickly growing threat to worldly biodiversity is light pollution [1]. The inappropriate exposure to lighting due to artificial light at night (ALAN) negatively influences the circadian system, inducing acute effects on sleep and cognition, as well as chronic endocrine-disrupting effects resulting in lifestyle diseases like obesity, cardiovascular disease, diabetes, and cancer [2]. The treatment with different lighting conditions can, at least in some cases, hold the circadian clock, and considerably reduce the sensitivity of rhythms towards certain drugs [3]. Despite the serious repercussions, the effect of ALAN at the transcriptomic level is yet to be studied deeply, especially because biological rhythms are also controlled transcriptionally. Here in this study, we have set up three experimental conditions, female zebrafish exposed to continuous light for one week, LLW, one month, LLM, and for one year, LLY, which revealed a clear desynchronization of circadian related genes as well as other genes in comparison to the normal 12-hour light and 12-hour dark, LD sample. The whole transcriptome combined data analysis of all four groups revealed 2309 genes important for healthy lifestyle are significantly up or down regulated. Further, a sample-to-sample comparison was also done to confirm the expression of genes in different levels of light entrapment. The development of several disease classes covering maximum available lifestyle disease types, and a novel gene signature with 28 unique genes (LDvsLLW: 14, LDvsLLM: 10, LDvsLLY: 12) reveals the effect of continuous light in zebrafish. We believe this result could help with the prognosis of disease classes including a wide array of neoplasms, urogenital diseases, pregnancy complications, mental disorders, endocrine system diseases, skin and connective tissue diseases, to name a few in patients.

References

[1] Koen EL, Minnaar C et al. Emerging threat of the 21st century lightscape to global biodiversity. *Global Change Biology.* 2018 Jun;24(6):2315-2324. doi: 10.1111/gcb.14146.

[2] Khan ZA, Labala RK et al. Artificial Light at Night (ALAN), an alarm to ovarian physiology: A study of possible chronodisruption on zebrafish (Danio rerio). *Science of The Total Environment.* 2018; 628-629: 1407-21.

[3] Falcón J, Torriglia A et al. Exposure to Artificial Light at Night and the Consequences for Flora, Fauna, and Ecosystems. *Frontiers in neuroscience.* 2020; vol. 14 602796. doi:10.3389/fnins.2020.602796

P-23

scdrake: a reproducible and scalable pipeline for scRNA-seq data analysis

Novotný J.[1,2], Kubovčiak J.[1], Kolář M.[1,2]

[1] Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 142 20 Prague 4, Czech Republic
[2] Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology in Prague, Technická 5, 166 28 Prague 6, Czech Republic

Motivation: While the workflow for primary analysis of single-cell RNA-seq data is well established, the secondary analysis of the feature-barcode matrix is usually done by custom scripts. There is no fully automated pipeline in the R statistical environment, which would follow the current best programming practices and requirements for reproducibility.
Results: We have developed scdrake, a fully automated workflow for secondary analysis of scRNA-seq data, which is fully implemented in the R language and built within the drake framework. The pipeline includes quality control, cell and gene filtering, normalization, detection of highly variable genes, dimensionality reduction, clustering, cell type annotation, detection of marker genes, differential expression analysis, and integration of multiple samples. The pipeline is reproducible and scalable, has an efficient execution, provides easy extendability and access to intermediate results, and outputs rich HTML reports.
Availability: The source code and documentation are available under the MIT license at https://github.com/bioinfocz/scdrake and https://bioinfocz.github.io/scdrake, respectively.

P-25

Exosomes produced by melanoma cells significantly influence the biological properties of normal and cancer-associated fibroblasts

Pfeiferová L.[1,2], Strnadová K.[3,4], Přikryl P.[5], Dvořánková B.[3,4], Vlčák E.[6], Frýdlová J.[5], Vokurka M.[5], Novotný J.[1,2], Šáchová J.[1], Hradilová M.[1], Brábek J.[7], Šmigová J.[7], Rösel D.[7], Smetana K. Jr.[3,4], Kolář M.[1,2], Lacina L.[3,4,8]

[1] Department of Informatics and Chemistry, University of Chemistry and Technology, Prague, Czech Republic
[2] Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[3] Institute of Anatomy, 1st Faculty of Medicine, Charles University, Prague 2, Czech Republic
[4] BIOCEV, 1st Faculty of Medicine, Charles University, Vestec, Czech Republic
[5] Institute of Pathological Physiology, 1st Faculty of Medicine, Prague, Charles University
[6] Electron Microscopy Core Facility, Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic
[7] BIOCEV, Faculty of Sciences, Charles University, Vestec, Czech Republic
[8] Department of Dermatovenereology, 1st Faculty of Medicine, Charles University and General, University Hospital, Prague, Czech Republic

The incidence of cutaneous malignant melanoma is increasing worldwide. While the treatment of the initial stages of the disease is simple, the advanced disease frequently remains fatal despite novel therapeutic options. This urges for identification of novel therapeutic targets in melanoma. Similar to other types of tumors, the cancer microenvironment plays a prominent role and determines the biological properties of melanoma. Importantly, melanoma cell-produced exosomes represent an important tool of intercellular communication within this cancer ecosystem. We have focused on potential differences in the activity of exosomes produced by melanoma cells towards melanoma-associated fibroblasts and normal dermal fibroblasts. Cancer-associated fibroblasts were activated by the melanoma cell-produced exosomes significantly more than their normal counterparts, as assessed by increased transcription of genes for inflammation-supporting cytokines and chemokines, namely IL-6 or IL-8. We have observed that the response is dependent on the duration of the stimulus via exosomes and also on the quantity of exosomes. Our study demonstrates that melanoma-produced exosomes significantly stimulate the tumor-promoting proinflammatory activity of cancer-associated fibroblasts. This may represent a potential new target of oncologic therapy.

P-27

QM-like partial atomic charges for AlphaFold available online

Schindler O.[1,2], Raček T.[1,2], Jelínek K.[1], Svobodová R.[1,2]

[1] *National Centre for Biomolecular Research, Faculty of Science, Masaryk University Brno, CZ*
[2] *CEITEC – Central European Institute of Technology, Masaryk University Brno, CZ*

Proteins are the basic functional unit of all living organisms. Critical information for understanding the function of a protein is its structure. Thanks to the AlphaFold algorithm [1], which predicts structure based on sequence, the number of predicted structures is growing very rapidly. Unfortunately, due to the computational complexity, we are unable to directly calculate the second key characteristic, i.e., electron density, for such large structures. A suitable approximation is the concept of partial atomic charges, which describe how much electron density belongs to each protein atom. Partial atomic charges can be derived directly from the electron density or might be calculated by fast empirical methods. However, these methods must go through a parameterization process, during which the parameters of an empirical method are optimized to reproduce the charges from quantum mechanics (QM).
This work introduces an empirical method called Split-charge equilibration with parameterized initial charges (SQE+qp) [2] adapted for AlphaFold Protein Structure Database. Our method can reproduce QM partial atomic charges with high accuracy. We also present an implementation of SQE+qp and its parameters via a web application Atomic Charge Calculator II [3] at https://acc2.ncbr.muni.cz. Thus, we provide the scientific community with a freely available online tool for calculating QM-like partial atomic charges.

References

[1] Varadi, M *et al*. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* (2021).
[2] Schindler, O *et al.* Optimized SQE atomic charges for peptides accessible via a web application. *Journal of Cheminformatics* (2021).
[3] Raček, T *et al.* Atomic Charge Calculator II: web-based tool for the calculation of partial atomic charges. *Nucleic Acids Research* (2020).

P-29

Measures of quality of clusters in hierarchical clustering of flow cytometry data

Sieger T.[1], Podolská T.[2], Fišer K.[2,3]

[1] Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic
[2] CLIP - Childhood Leukaemia Investigation Prague, Department of Paediatric Haematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic
[3] Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic

Hierarchical clustering enables unsupervised analysis of multidimensional data, yielding a dendrogram, a hierarchical tree of clusters of data samples. However, the dendrogram does not readily specify the quality of individual clusters in it. If users need to choose „good“ clusters out of the dendrogram, traditionally, they cut the dendrogram at a specific height and pick top-level clusters, or manually cherry pick some clusters. Unfortunately, this would miss „good“ clusters appearing at different heights in the dendrogram, and could be subjective. We attempted to find measures of quality of individual clusters in hierarchical clustering that could be used to guide selection of clusters in flow cytometry data both in manual and automated fashion. We defined theoretical requirements of two measures of quality of clusters in a dendrogram: the compactness, assessing how tightly each cluster is connected to the hierarchy below it, and the separation, assessing how well each cluster is separated from the hierarchy above it. Notably, these quality measures do not intentionally rely on original data, but purely on the dendrogram itself. We devised and implemented nontrivial measures fulfilling the requirements mentioned above and validated them on flow cytometry data (n=10) with annotation of „good“ clusters from two independent researchers. The median intra-rater classification accuracy was 95.0%, and the median inter-rater accuracy was 92.2%. We confirmed that clusters with high values of the theoretically derived measures of quality often corresponded to „good“ ground truth clusters, and that the quality measures enabled to select clusters from the dendrogram automatically with the accuracy of 88.9%. We devised measures of quality of clusters in hierarchical clustering that can be used to guide selection of meaningful clusters in flow cytometry data. Our unsupervised method enables automated processing of large amounts of data without a need of costly and subjective manual intervention. In future, we need to study the general applicability of our measures of cluster quality and validate them on other data sets.

This work was supported by GAUK number 352922.

P-31

Combination of expert decision systems with artificial intelligence leads to superior accuracy of automated prediction of clinical effect of copy number variation

Sládeček T.[1,5], Gažiová M.[1,2], Pös O.[1,5], Pös Z.[1,3,4], Budiš J.[1,5,6], Radvánszky J.[1,4,5], Szemes T.[1,3,5]

[1] Geneton s.r.o., Bratislava 84104, Slovakia
[2] Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava 84248, Slovakia
[3] Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava 84215, Slovakia
[4] Institute of Clinical and Translational Research, Biomedical Research Center, Slovak Academy of Sciences, Bratislava 84505, Slovakia
[5] Comenius University Science Park, Bratislava 84104, Slovakia
[6] Slovak Center of Scientific and Technical Information, Bratislava 81104, Slovakia

Interpretation of clinical impact of large mutations is a difficult task mainly due to the size of affected genomic content. In the past years, several tools have been built for interpretation of single nucleotide polymorphisms (SNP), however not so many for copy number variants (CNV). In this work we present two methods for pathogenicity prediction of CNVs: an ACMG based method which follows strict set of rules aggreed upon by a large consortium of scientists and a machine-learning based method ISV (Interpretation of Structural Variation). We show that a joint use of both tools yields superior performance compared to using the tools alone. We believe that using predictors utilizing different data sources should create a more robust overal predictor, which can be used in practice by laboratory scientists to help with the laborious interpretation process.

P-33

PredictSNP^{ONCO}: A Web Server for Rapid Structural Bioinformatics Analysis of the Effect of Cancer-associated Mutations

Stourac J.[1,2], Khan Rayyan T.[1,2], Borko S.[1], Pokorna P.[3,4], Dobias A.[1], Pinto G.[1,2], Sterba J.[5,6], Slaby O.[3,4], Bednar D.[1,2], Damborsky J.[1,2]

[1] *Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Masaryk University, Kamenice 5, 625 00 Brno*
[2] *International Clinical Research Centre, St. Anne's University Hospital Brno, Pekařská 53, 656 91 Brno*
[3] *Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno*
[4] *Department of Biology, Faculty of Medicine, Masaryk University, Kamenice 5, 625 00 Brno*
[5] *Department of Paediatric Oncology, University Hospital Brno, Černopolní 212, 613 00 Brno*
[6] *Faculty of Medicine, Masaryk University, Kamenice 5, 625 00 Brno*

Cancer is one of the top leading causes of death in the world. Due to its complex nature, there is not a single general-purpose treatment, and the existing ones have varying success rates for different types of cancer. Thus, there is a massive demand for alternative approaches capable of delivering the proper treatment at the right time. One of the most promising approaches is precision medicine which breaks the traditional "one size fits all" paradigm by considering each patient as a unique individual and designing their treatment plan according to their needs. Precision medicine exploits data collected from the patients, such as mutations detected in the affected tissues and their anamnesis. Most of the pipelines used in clinical practice are based on analysing the mutation impact on the DNA level. As many mutations can also occur in the exonic regions of the DNA, this analysis can lack essential data about their effect on the protein level. To fill this gap, we developed a novel web server PredictSNP^{ONCO} capable of rapid *in silico* assessment of the mutation effect on multiple essential properties of proteins such as function and stability. The server provides various structure-based analyses such as stability evaluation, virtual screening of potential inhibitors using the dataset of FDA/EMA approved drugs, as well as sequence-based analyses like conservation analysis and prediction of mutational effect on protein function. Important annotations extracted from state-of-the-art databases complement calculated values. The service offers an easy-to-use interactive web interface that allows users to start the analysis and evaluate the results efficiently. The server is available freely to the scientific and medical community at https://loschmidt.chemi.muni.cz/predictsnp-onco.

P-35

Analysis of sequencing data from reprogramming of immortalized cell line

Svatoňová P.[1]

[1] *Institute of Molecular Genetics*

For exploring erythroid progenitors reprogramming of zebrafish immortalized cell line, bulk ATAC-Seq together with RNA-Seq was performed and analysed. After preprocessing part, where quality control (FastQC), trimming (Cutadapt/Trimmomatic), mapping (HISAT2/Salmon) and filtering (Picards MarkDuplicates, Samtools/SortMeRNA) were included, regions of interest were detected for ATAC-Seq via peak calling (Macs2). Then quantification (FeatureCounts/Salmon quant, Tximport) were performed, followed by differential analysis (DESeq2). Reprogramming of cell type involves changes on many levels. Indeed, thousands of genes were significantly deregulated, tens of thousands peaks marked differentially abundant regions on genome. To reduce these huge numbers and gain more biological view, genes in defined proximities from peaks were considered. Not surprisingly, hemoglobins with other heme/oxygen binding/carrying genes occur among top10 downregulated genes. In contrast, most upregulated genes are markers of myeloid cells. That supports observed changes and proves successful reprogramming.

P-37

Bioinformatic pipeline for comprehensive analysis of various small RNAs through RNA sequencing

Trachtova K.[1,2], Bystry V.[1], Demko M.[1], Slaby O.[1,3]

[1] *Central European Institute of Technology, Masaryk University, 60177 Brno, Czech Republic*
[2] *Christian Doppler Laboratory for Applied Metabolomics (CDL-AM), Medical University of Vienna, 1090 Vienna, Austria*
[3] *Department of Biology, Faculty of Medicine, Masaryk University, 60177 Brno, Czech Republic*

Next-generation sequencing (NGS) is a revolutionary method that allows massive parallel sequencing of millions of DNA or RNA fragments. Although NGS is considered a state-of-the-art method, there is still a need for more comprehensive bioinformatical approaches, especially in the research of various small RNAs. In the sequencing of small RNAs, the crucial problem is accurate identification and quantification of a full spectrum of small RNA pool. Most of the available workflows are however targeted mostly at microRNA and ignore other RNA types such as snoRNA, snRNA, piRNA, or isomiRs.
We propose here a bioinformatic pipeline for accurate quantification of all known small RNAs classes. Our pipeline is divided into stand-alone modules, each focusing on one part of the sequencing data analysis (first quality control, pre-processing, RNA quantification and differential expression analysis). The most significant is the RNA quantification module, where a subsequent number of mapping rounds, utilizing reference sequences collected from several resources, ensure quantification of all different small non-coding RNAs. Custom Python tool was created to count reads assigned to different RNAs that also address an issue of multi-loci RNAs (such as piRNA) and problem of overlapping RNA annotations.
Each module provides a PDF/HTML report summarizing results, including tables, plots and their explanation and so guiding the user when exploring different small RNA expression levels. To smooth utilization of report plots for publication, we also offer an interactive application implemented in Shiny for a real-time visualization of differential expression results where the content and appearance of popular plots such as heatmap, PCA or volcano plot can be easily altered.

Core Facility Bioinformatics of CEITEC Masaryk University is gratefully acknowledged for the obtaining of the scientific data presented here.

Poster session

Tuesday, 14. June

Poster session is sponsored by the
National Infrastructure of Chemical Biology.

CZ OPENSCREEN
National Infrastructure for Chemical Biology

P-2

Detection of tail fibre proteins via machine learning methods

Bača J.[1], Baláž A.[1,2]

[1] *Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*
[2] *Geneton Ltd., Bratislava, Slovakia*

Bacteriophages are viral organisms infecting and killing bacteria. They contain tail fibre proteins, which can be utilised in the diagnostics and consequent treatment of bacterial infections. The identification of these proteins in the laboratory is expensive and time-consuming. In this work, we present a bioinformatics tool, capable of detecting tail fibre proteins from a set of proteins with unknown functions. To create this tool, we utilised the information from public databases, identified common features of tail fibre proteins, and used those features to train multiple machine learning models. The models were evaluated and the best performing model was deployed to the Anaconda cloud for the ease of installation and use. The comparison of the model's performance with current state-of-art models on the test set showed significant improvement in F1 score for the tail fibre detection.

P-4

Bioinformatics workflow for reliable detection of SARS-CoV-2 variants in wastewater data generated by massively parallel sequencing

Böhmer M.[1,2], Čárska D.[3], Hekel R.[1,4,5], Sládeček T.[1], Sitarčík J.[1,3,5], Goga A.[1,3], Krampl W.[1,4], Rusňáková D.[1,2,4], Sedláčková T.[1], Kaliňáková A.[2], Budiš J.[1,5], Szemes T.[1,2,4,5]

[1] *Comenius University Science Park, Bratislava, Slovakia*
[2] *Public Health Authority of the Slovak Republic, 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] *Slovak Centre of Scientific and Technical Information, Bratislava, Slovakia*

The ongoing SARS-CoV-2 pandemic has caused the deaths of more than 6.2 million people around the world according to the World Health Organisation. It has urged us to find a way to predict emerging hotspots and therefore, with good precautionary measures, reduce the possibility of new cases and following deaths. Monitoring of wastewaters with the inclusion of massively parallel sequencing of wastewater samples appears to be an excellent tool for this purpose. In fact, novel SARS-CoV-2 variants from wastewater could be identified 1 to 2 weeks before being detected in clinical samples from the same area. The problem arises with analysing such data. In this work, we present a highly scalable and easily deployable computational pipeline for reliable detection of SARS-CoV-2 variants of concern in wastewater data generated by massively parallel sequencing. It uses state-of-the-art tools freely available online combined with our developed method to distinguish between SARS-CoV-2 variants of concern. Our bioinformatics workflow is currently used to monitor wastewater for SARS-CoV-2 variants in the Slovak republic which is coordinated by the national Public health authority.

P-6

Mining novel terpene synthases from large-scale sequence repositories

Čalounová T.[1,2], Tajovská A.[1], Smrčková H.[1,2], Bushuiev R.[1,3], Samusevich R.[1,4,5], Pluskal T.[1]

[1] *Institute of Organic Chemistry and Biochemistry CAS*
[2] *Faculty of Science, Charles University in Prague*
[3] *Faculty of Information Technology, Czech Technical University in Prague*
[4] *Czech Institute of Informatics, Robotics and Cybernetics (CIIRC CTU)*
[5] *Faculty of Chemical Technology, University of Chemistry and Technology, Prague*

Terpenoids are the largest and most diverse group of plant specialized metabolites with numerous medical or industrial applications. The terpenoid scaffolds are produced by enzymes called terpene synthases, which catalyze some of the most complex chemical reactions in biology.

We have created a database of characterized terpene synthases, which to this date contains more than one thousand entries, each including information about their taxonomy, terpene types and reactions. These terpene synthases were subjected to analysis of their domain architectures and structures, especially with respect to their taxonomy and terpene types. To perform the analysis of structures, we used AlphaFold 2 to obtain structure predictions for all terpene synthases in the database since there are very few experimentally determined structures of terpene synthases.

We mined UniParc, 1KP, Phytozome and TSA protein databases (606,214,611 protein sequences) using hidden Markov models from the Pfam database, which are associated with terpene synthases. From these databases we obtained 191,476 protein sequences of putative terpene synthases. We further generated a phylogenetic tree from both characterized and mined terpene synthase sequences. We have analyzed the phylogenetic tree and selected promising uncharacterized candidates based on the largest phylogenetic distance to the characterized sequences in the tree and their sequence reliability scores. The reliability score was calculated using the knowledge gained from the analysis of the database of characterized terpene synthases, primarily focusing on their sequence lengths and domain architectures.

Several selected candidate sequences will be experimentally characterized. Since we prioritize sequences distant from already characterized sequences, we presume that these sequences could synthesize novel terpene scaffolds.

P-8

ExP Heatmap: visualization of high-dimensional pairwise genomic data

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

[1] *Laboratory of Genomics and Bioinformatics, Institute of Molecular Genetics of the ASCR, v. v. i., Vídeňská 1083, 142 20 Prague 4, Czech Republic*
[*] *edvard.ehler@img.cas.cz*

The Cross-Population Heatmap (ExP heatmap) is a novel method developed by our team to display highly dimensional cross-population data. It is suitable to display several thousands of data-points, each with hundreds of dimensions. We are aiming this tool to cross-population (or pair-wise) data, primarily on 1000 Genomes Project, phase 3, genomic data, but the method can be easily extended or adapted for usage on essentially any type of data that are in form of similarities/distances between groups. The most profound advantage of our method is the ability to display several millions of results (i.e. p-values, distances) in one picture, while allowing the user to clearly identify significant patterns or important genome areas by his or her own sight. All this with fast and user-friendly implementation.

The ExP method was implemented in Python 3.8+ and is ready to be used as Python package or stand-alone command-line tool. It is available from official Python Package Index (PyPI, https://pypi.org/project/exp-selection/1.0.0/) or GitHub (https://github.com/ondra-m/exp-selection). We are looking forward to several updates of ExP heatmap software coming up in our pipeline.

P-10

Formalized machine-assisted flow cytometry immunophenotyping

Sieger T.[1], Koblížek M.[2], Fišer K.[3,4]

[1] *Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Prague, Czech Republic*
[2] *Department of Pathology and Molecular Medicine, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic*
[3] *CLIP-Childhood Leukaemia Investigation Prague, Department of Pediatric Hematology and Oncology, Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic*
[4] *Department of Bioinformatics, Second Faculty of Medicine, Charles University, Prague, Czech Republic*

Flow cytometry is a method for quantitative measuring of multiple parameters on millions of individual cells. Such ability is widely used both in research and diagnostics. For example, in hemato-oncology, the identification of malignant cell population by flow cytometry is a part of diagnostics and disease monitoring. Apart from idenitification of cell population its description, so call immunophenotype, is important for clinical decision making.

However the guidelines for description of the immunophenotype vary between laboratories despite multiple efforts for standardization. It is mainly due to the fact that the descriptions are formally imprecise and so leave space for subjective interpretations.

Here we present a system of deriving immunophenotypes automatically from data with labeled cells of interest and cells serving as negative reference. First we formalized several guidelines described in the literature. However, more importantly we also formalized immunophenotyping further by providing continuous measures of cell population positivities for individual analysed parameters in terms of distances between population of interest and reference negative cell population. We first defined a set of formal criteria (n = 9) for such measures. Then we considered known distance measures (n = 7) including Separation Index, Kullback-Leibler divergence and Marker Enrichment Modeling score, and added our own measures (n = 13) based mostly on ROC-AUC and Earth movers distances, and tested all measures for fulfilling the previously set criteria. To test the performance of selected measures we used synthetic data sampled from known distributions and selected best performing measures. To be able to compare the best performing measure to expert immunophenotype, the output of the measure was discretized into three categories (negative, medium and positive) . When used on two real world data sets, such categories were in agreement with expert-generated categories in 27 out of 28 parameters used to define 7 cell populations. As a last step we created a framework to automatically generate pdf or html reports from provided labeled data.

Overall we provide: 1) formalized versions of immunophenotype reporting rules, 2) a set of measures of population positivity based on distribution distances and evaluation of their peformance, and 3) R-based tools to generate immunophenotype report automatically from labeled data.

This work was supported by UNCE/MED/015.

P-12

miRBind: a Deep Learning method for miRNA binding classification

Klimentova E.[1], Hejret V.[2], Giassa I.[2], Alexiou P.[2]

[1] *Faculty of Informatics, Masaryk University, Brno, Czechia*
[2] *Central European Institute of Technology (CEITEC), Masaryk University, Brno, Czechia*

The binding of microRNAs (miRNAs) to their target sites is a complex process, mediated by the Argonaute (Ago) family of proteins. Prediction of miRNA:target site binding is an important first step for any miRNA target prediction algorithm. To date, the potential for miRNA:target site binding is evaluated either using co-folding free energy measures, or heuristic approaches based on the identification of binding 'seeds', i.e. continuous stretches of binding corresponding to specific parts of the miRNA. The limitations of both these families of methods have produced generations of miRNA target prediction algorithms exclusively focused on 'canonical' seed targets, even though unbiased experimental methods have shown that only approximately half of in vivo miRNA targets are 'canonical'. Here we present miRBind, a deep learning method and web-server that can be used to accurately predict the potential of miRNA:target site binding. We train our method on seed-agnostic experimental data, and show that our method outperforms both seed-based approaches and co-fold free energy approaches. The full code for development of miRBind is freely available at https://github.com/ML-Bioinfo-CEITEC/miRBind, and a free web-server is available at https://ml-bioinfo-ceitec.github.io/miRBind

P-14

Diversity and transcriptome analysis of the maize B chromosome

Hloušková L.[1*], Holušová K.[1], Karafiátová M.[1], Bartoš J.[1]

[1] *Institute of Experimental Botany of the Czech Academy of Sciences, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc 77900, Czech Republic*
[*] *simkoval@ueb.cas.cz*

B chromosomes are enigmatic elements in thousands of plant and animal genomes that persist in populations despite being nonessential. Maize (Zea mays L.) is one of the most important crops and serves as well established model for biological research and the maize B chromosome has been studied for many decades. However, the diversity of the maize B chromosome and its distribution within maize landraces are still not thoroughly described as well as its gene expression. Reference sequence of the maize B chromosome revealed 758 protein coding genes (Blavet et al., 2021). Here we present the first results of a gene expression atlas of 12 maize tissues using RNA-seq and gene expression pipeline based on mapping reads to the reference genome of Z. mays (B73 RefGen_v4) supplemented with the reference of the maize B chromosome (Zm-B73_B_CHROMOSOME_MBSC-1.0, Blavet et al., 2021) and subsequent differential-expression (DE) and gene ontology (GO) enrichment analyses.

Funding: Ministry of Education, Youth and Sports; Collaboration with CIMMYT on the study of diversity and evolution of maize B chromosome (LTT19007)

References

[1] Blavet et al. (2021) Sequence of the supernumerary B chromosome of maize provides insight into its drive mechanism and evolution. PNAS 118 (23): e2104254118

P-16

NanoLuc luciferase may not be as “nano” as thought

Horáčková J.[1,2], Nemergut M.[1,2], Pluskal D.[1,2], Marques S.[1,2,3], Blechová V.[1,2,3], Tulis J.[1,2], Damborský J.[1,2,3], Prokop Z.[1,2,3], Janin Y.[4], Marek M.[1,2,3], Bednář D.[1,2,3]

[1] Loschmidt Laboratories, Department of Experimental Biology, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
[2] Loschmidt Laboratories, RECETOX, Faculty of Science, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
[3] International Clinical Research Centre, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic
[4] Structure et Instabilité des Génomes (StrInG), Muséum National d'Histoire Naturelle, INSERM, CNRS, Alliance Sorbonne Université, 75005 Paris, France

NanoLuc is a recently developed commercially available bioluminescent enzyme for broad biotechnological and biomedical applications. One of the important declared advantages of NanoLuc is its small size of 171 amino acid residues compared to conventional luciferases from firefly (550 residues) and *Renilla reniformis* (312 residues). However, the experiments with NanoLuc conducted in our laboratory suggested that NanoLuc may homodimerize during the catalytic cycle, and thus the catalytically active bioluminescent system would be twice as large. Here we present a computational study employing molecular docking and enhanced sampling methods for molecular simulations to study substrate binding, protein dynamics, and dimer dissociation. These findings support the lab-based experiments and provide a bigger picture of the function of this widely used enzyme. Our findings suggest that while NanoLuc is a monomer in a substrate-free solution, in the presence of its substrate it is most likely a dimer, which should be considered when designing experiments with NanoLuc.

P-18

Comparison of characteristics of cfDNA fragments between patients with severe SARS-COV-2 infection and healthy control group

Krampl W.[1,2,5], Rusňáková D.[1,2,3,5], Sedláčková T.[1,2], Hodosy J.[1,5], Böhmer M.[1,2,3], Budiš J.[1,2,6], Szemes T.[1,2,5]

[1] *Comenius University Science Park, Bratislava, Slovakia*
[2] *Geneton Ltd., Bratislava, Slovakia*
[3] *Public Health Authority of the Slovak Republic, Bratislava, Slovakia*
[4] *Institute of Molecular Biomedicine, Faculty of medicine, Comenius University, Bratislava, Slovakia*
[5] *Department of Molecular Biology, Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia*
[6] *Slovak Center of Scientific and Technical Information, Bratislava, Slovakia*

SARS-COV-2 virus (Severe acute respiratory syndrome coronavirus 2) is a human pathogen causing a potentially fatal disease COVID-19, responsible for a world pandemic. Progress of COVID-19 in an infected patient might be asymptomatic, causing only light symptoms or severe symptoms that might result in death of the patient (mortality rate 1.21%). The mechanism behind the severity of a patient's symptoms is currently unknown - with the exception of statistically inferred groups of people being at a higher risk of developing severe symptoms such as seniors, immunosuppressed patients or patients suffering from asthma. In our work, we have compared characteristics of cfDNA fragments isolated from blood plasma of patients who have developed severe COVID-19 symptoms with healthy control group's cfDNA fragments. One of several characteristics with statistically significant difference is fragment length with patients's fragments (median size 99.5) being longer in comparison to control group's fragments (median size 81).

P-20

PrankWeb3 - binding site predictions for experimental and predicted protein structures

Jakubec D.[1], Škoda P.[1], Krivak R.[1], Novotný M.[2], Hoksza D.[1]

[1] *Faculty of Mathematics and Physics, Charles University*
[2] *Faculty of Science, Charles University*

PrankWeb is a web-based state-of-the-art ligand-binding site prediction tool. We introduce a new version with two major and an array of minor improvements. The major improvements involve a new, faster and more accurate evolutionary conservation estimation pipeline and the ability to carry out LBS predictions in situations where no experimental structure is available.

The original version of PrankWeb utilized an evolutionary conservation calculation pipeline which attempted to reproduce the evolutionary history of the queried sequence and could take up to several hours to finish. In PrankWeb 3.0, we replaced the evolutionary rate-based conservation scores with an entropy-based metric. The new conservation calculation pipeline utilizes the HMMER3 package for fast and sensitive sequence similarity searches against the UniRef50 sequence database. This change led to more accurate predictions and enabled us to reduce the average time required for the conservation score calculations down to a few minutes.

To extend the functionality of PrankWeb to proteins with no experimental structures available, we have pre-computed the LBS predictions for the structural models from the newly developed AlphaFold database (ADB) and integrated the results into our web server. The user has now the option to enter a UniProt accession number; then, if available, the corresponding model is fetched from the ADB, LBSs are predicted and presented to the user. This required training a new machine learning model specialized for the AlphaFold structures and adapting the interface to visualize specific features such as residue-level confidence scores.

The minor improvements include the ability to deploy PrankWeb as a Docker container, support for the mmCIF file format, improved public REST API access, or the ability to batch download the LBS predictions for the whole PDB archive and parts of the ADB.

P-22

HCV IRES Secondary Structure Search in Human 5'UTRs

Le Anh Vu[1]

[1] *Czech Technical University*

Cap-independent translation is a well-known mechanism that viral RNAs use to promote their transcription at the expense of cellular mRNAs. Though typical for viruses, a fraction of human genes has been shown in vitro to use this mechanism as well. Key to this process is a so-called internal ribosomal entry site (IRES) - an RNA element able to recruit ribosomes without the canonical set of transcription factors. The function of IRESs is closely tied to their structure and until now, 4 major types of IRESs have been described, differing in the structural organization. Type IV is also known as HCV-like IRES - named after the notorious human pathogen - Hepatitis C virus (HCV). This work hypothesizes, that a structure similar to the one of HCV IRES can be found in human 5'UTR sequences and proposes a computational pipeline that outputs potential candidates resembling the target structure. The keystones of the pipeline are programs RSEARCH and NA2Dsearch. RSEARCH provides a fast and broad collection of preliminary matches, NA2DSearch subsequently filters the
matches via folding and structural matching. We present 118 findings in 5'UTR regions that contain structural motifs, which in silico may exhibit transcription regulatory capabilities of HCV IRES.

P-24

Investigating the Interaction Between Human DHX15 Helicase and Viral G-Patch Domain Within the Virions of Mason-Pfizer Monkey Virus Using UV Crosslinking and Immunoprecipitation

Pavlů A.[1], Dostálková A.[2], Křížová I.[2], Kolář M.[3], Rumlová M.[2], Ruml T.[1]

[1] *Department of Biochemistry and Microbiology, University of Chemical Technology, Prague, Czech Republic*
[2] *Department of Biotechnology, University of Chemical Technology, Prague, Czech Republic*
[3] *Institute of Molecular Genetics, Czech Academy of Sciences, Prague, Czech Republic*

The DHX15 helicase is a member of the DEAH/RHA helicase family. It is known to take part in several critical biochemical pathways including splicing of pre-mRNA or ribosome biogenesis. A subset of this helicase family is known to be regulated by a group of proteins which contain a conserved glycine-rich motif called G-patch domain (GPD). Moreover, the GPD has been proposed as an interaction intermediary between the helicase and the GPD-containing protein, aiding in the recruitment. Interestingly, this almost exclusively eukaryotic motif was also detected in the genome of several members of the *Betaretroviridae* family, including Mason-Pfizer monkey virus (M-PMV). M-PMV is a simple retrovirus used often as a model to study the life cycle of retroviruses and the GPD coding sequence is localized at the 3' end of the *pro* gene, just upstream of the site of the second ribosomal shift. The GPD is a part of viral protease; however, it has been observed that in some cases it could be part of the M-PMV reverse transcriptase (RT). Taken together, we aimed to investigate whether the RT (through the GPD) interacts with the human DHX15, with viral genomic RNA being the site of the potential interaction. We used the crosslinking and immunoprecipitation coupled with sequencing (CLIP-seq) approach. Both DHX15 and GPD were immunoprecipitated independently from mature M-PMV virions and the RNA bound to these proteins was isolated, purified and sequenced. We postulated a hypothesis that if the two proteins with RNA-binding capabilities do interact with each other we would observe an overlap of detected binding sites. Indeed, our data suggest that these overlaps are present and thus provide support for the potential interaction between the human DHX15 and RT of the M-PMV in mature virions.

P-26

Atomic Charge Calculator II – a web service for calculating partial atomic charges

Raček T.[1,2], Schindler O.[1,2], Ptáčníková L.[2], Svobodová R.[1,2]

[1] *CEITEC Masaryk University, Kamenice 753/5, Brno, Czech Republic*
[2] *National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 753/5, Brno, Czech Republic*

Partial atomic charges are a theoretical concept simplifying the electron density in a molecule to a set of point charges. Charges were proven useful in many research areas, e.g., computational chemistry or structural bioinformatics. Since they are not physical observables, their values must be computed via a suitable method. Standard approaches employ quantum mechanics (QM) principles, but their high computational complexity limits their use to small systems. On the other hand, empirical methods can reproduce the results from the QM calculations with high accuracy within a fraction of the original time. Unfortunately, implementations of most empirical methods are not available to the users. Therefore, we developed a web service Atomic Charge Calculator II (ACC II) [1], to make the most important empirical methods accessible to the community. ACC II features 20 empirical methods with parameter sets gathered from the literature. It can be used interactively in a web browser providing charge visualizations, or automatically in user-defined workflows utilizing an API. ACC II is freely available at https://acc2.ncbr.muni.cz.

References

[1] Raček, T., Schindler, O., Toušek, D., Horský, V., Berka, K., Koča, J., & Svobodová, R. (2020). Atomic Charge Calculator II: web-based tool for the calculation of partial atomic charges. Nucleic acids research, 48(W1), W591-W596.

P-28

Exploring possibility of reusing LIGR-seq datasets for biological data supported RNA secondary structure prediction

Schwarz M.[1]

[1] *Institute of Microbiology of the Czech Academy of Sciences*

LIGR-seq and PARIS are methods for studying RNA-RNA interactions *in vivo* based on reversible crosslinking of double-stranded RNA mediated by a psolaren derivative AMT. Albeit names the methods are very different, the methods themselves are very similar, briefly: crosslinking of double-stranded RNA *in vivo*, digestion of single-stranded RNA, proximity ligation, decrosslinking and sequencing. Apart from the primary objective of detecting intermolecular RNA-RNA interactions, the data from PARIS method were also used to detect intramolecular interactions, thus providing evidence of based paired regions from single RNA molecule. These can be used for directed secondary structure prediction with IRIS method which was developed for PARIS data. Here we explore if there are sequencing reads supporting intramolecular interactions present in *Bacillus subtilis* LIGR-seq dataset and the possibility to use the IRIS method for the secondary structure prediction.

P-30

Probes & Drugs Portal in 2022: A Hub for the Integration of High-quality Bioactive Compound Sets

Škuta C.[1], Bartůněk P.[1]

[1] *CZ-OPENSCREEN National Infrastructure for Chemical Biology, Institute of Molecular Genetics of the Czech Academy of Sciences, Vídeňská 1083, 142 20, Prague 4, Czech Republic*

Probes & Drugs (P&D) portal (probes-drugs.org) was released in 2017 to address the problems in the field of chemical probes and other high-quality chemical tools, concerning mainly the fragmentation and obsoleteness of the available data. Since then, it has become one of the major and most comprehensive resources in the field that serves as a hub for the integration of high-quality bioactive compound sets. Apart from the data integration, P&D recently established its own approach to evaluating the probe-likeness of compounds labelled as chemical probes (so-called probe-likeness score) and a live set of high-quality chemical probes (utilizing the score in combination with other key criteria), updated with each new version of P&D. As of ver. 01.2022, this set contains 669 compounds covering 518 primary targets. The synergy between integrated data sources and tools available for their analysis makes P&D a unique discovery platform designated not only for experts in the field of chemical biology but for a wider scientific community with limited knowledge in this area.

P-32

Exploring the possibilities of formulation the rules for permeation enhancement

Storchmannová K.[1], Balouch M.[2], Štěpánek F.[2], Berka K.[1]

[1] Department of Physical Chemistry, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic.
[2] Department of Chemical Engineering, University of Chemistry and Technology, Prague, Technická 3, 166 28 Prague 6, Czech Republic.

In 1995, the FDA approved the first application of liposomes as lipid carriers of an active pharmaceutical ingredient (API) [1]. Since then, liposomes have been intensively studied. So far, there are more than 20 liposome formulations approved in the EU and USA, including COVID mRNA vaccines, and many others are undergoing clinical trials. Unfortunately, many APIs are hard or nearly impossible to formulate in liposomes because of their permeability and partitioning. These key properties of APIs were studied in our previous study [2]. There are few options for solving this problem, e.g. (i) a change of liposomes, (ii) adding a permeability enhancer or (iii) a small modification of the chemical structure of the API. We focused on the last option. We asked, "How much does the permeability of a small molecule change if we change its structure?". We analyzed the experimental permeability data from the MolMeDB database [3] from two experimental methods - CACO2 and PAMPA and one calculated method - COSMOPerm [4]. As a toy system, we studied the effect of structural changes on the permeability of cytarabine. We would like to present how the permeability of cytarabine changes if we systematically modify its structure.

References

[1] Barenholz, Y. (Chezy). Doxil® — The first FDA-approved nano-drug: Lessons learned. *J. Control. Release* **160**, 117–134 (2012).
[2] Balouch, M. *et al.* In silico screening of drug candidates for thermoresponsive liposome formulations. *Mol. Syst. Des. Eng.* **6**, 368–380 (2021).
[3] Juračka, J. *et al.* MolMeDB: Molecules on Membranes Database. *Database* **2019**, (2019).
[4] Schwöbel, J. A. H. *et al.* COSMO perm: Mechanistic Prediction of Passive Membrane Permeability for Neutral Compounds and Ions and Its pH Dependence. *J. Phys. Chem. B* **124**, 3343–3354 (2020).

P-34

NextDOM 2.0: Detector of Somatic Point Mutations in Leukemias Resistant to Therapy

Suchánková P.[1], Benešová A.[1], Pecherková P.[1,2], Polívková V.[1], Koblihová J.[1], Machová Polaková K.[1,3]

[1] Institute of Hematology and Blood Transfusion, Prague, Czech Republic
[2] Faculty of Transportation Sciences, Czech Technical University, Prague, Czech Republic
[3] Institute of Pathological Physiology, 1st Medicine Faculty, Charles University, Prague, Czech Republic

Chronic myeloid leukemia (CML) and Ph+ acute lymphoblastic leukemia (Ph+ ALL) are diseases associated with a characteristic chromosomal translocation called the Philadelphia chromosome. This chromosomal abnormality generates the fusion gene encoding BCR::ABL1 oncoprotein. The development of mutations in the kinase domain (KD) of BCR::ABL1 is a known mechanism leading to the resistance to CML and Ph+ ALL treatment, therefore early and correct mutation detection is crucial for better disease management.

Next-generation sequencing (NGS) of the targeted region of BCR::ABL1 KD enables to detect point mutations at very low-levels. In common practice, universal thresholds for mutation calling are used (most frequently 5% or 3%), which can lead to false negative or false positive results. Therefore, for precise variant calling and correct error detection, it is necessary to set individual thresholds for each sequenced position.

The NextDOM Set of Thresholds Creator calculates the individual threshold level for each sequenced position individually. This Set of Thresholds calculation uses NGS data from samples of healthy donors expecting no somatic point mutations. The appropriate thresholds are calculated only for positions that meet the criteria for a data quality control (number of pair reads). NextDOM applies this Set of Thresholds to NGS data of the patient sample and reports the mutations at statistically significant levels. The program also reports about clinical relevance of significant variants using the up-loaded list of the mutations which have been published in connection with resistance to therapy.

The NextDOM Set of Thresholds Creator can be used as a universal thresholds calculator and the NextDOM is modifiable for determining relevant mutation detection in various amplicon sequencing data.

P-36

Analysis of Molecular Simulation by Adversarial Autoencoder

Tedeschi G.[1], Višňovský V.[2], Křenek A.[2], Spiwok V.[1]

[1] Department of Biochemistry and Microbiology – University of Chemistry and Technology, Prague
[2] Department of Machine Learning and Data Processing – Faculty of Informatics - Masaryk University, Brno

Machine learning and artificial neural networks are intensively studied in connection with molecular simulations. This research is often motivated by acceleration. Molecular simulations, namely, the molecular dynamics simulation or Monte Carlo method, have a great potential in designing new drugs, proteins, enzymes, or materials. In principle, it is possible to simulate drug-target complexes, proteins or new materials to predict their stability and other properties from the evolution of molecular structure. However, the practical application of molecular simulations is complicated by their large computational costs. A typical biomolecular system consists of thousands of atoms interacting with each other via short-range and long-range non covalent interactions. This vast number of mutual interactions must be evaluated in every step of the simulation. Furthermore, a simulation step must be relatively short (femtoseconds in molecular dynamics) to assure numerical stability. As the results, typical molecular dynamics simulations can sample only a small fraction of the states available to the simulated system, with the likely loss of some slow or rarely occurring processes. There are numerous opportunities for machine learning and artificial neural networks to address this problem. For the reasons mentioned above, we apply autoencoders, generative neural networks and their combination as a platform for analysis of simulation data. The potential of this fusion was demonstrated on microsecond trajectory of Alanine Dipeptide and Trp-cage.

This work was supported by the Czech Science Foundation (22-29667S)

P-38

European Chemical Biology Database

Voršilák M.[1,2], Müller T.[1], Škuta C.[1], 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, 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, Technická 5, 166 28, Prague, Czech Republic

EU-OPENSCREEN (EU-OS) is a European research infrastructure for chemical biology founded in 2018 to support chemical probe and drug discovery projects. Currently, it integrates more than 20 high-throughput screening and chemistry facilities in eight different European countries. The European Chemical Biology Database (ECBD, https://ecbd.eu) is the central data hub for data generated within the EU-OS network. The ECBD is operated by the Institute Molecular Genetics of the Czech Academy of Sciences (IMG) in Prague. ECBD is a web portal with powerful search and analysis capabilities and contains validated output from screening centres in a public as well as pre-release environment. The ECBD is developed in line with the FAIR principles ensuring Findability, Accessibility, Interoperability, and Reusability of the data. The data are deposited with a flexible privacy model for rapid and safe dissemination and exploitation. Besides the web UI, ECBD also offers the data access through an API and a database dump with all public data. As of May 2022, there are 5 public data sets available to the wide scientific community with many more to be released once their embargo time period (up to 36 months) expires. To ensure high reliability and safety of the service, ECBD is hosted on the CESNET servers.

Al Tukmachi Dagmar *(dasa.altukmachi@gmail.com)*
CEITEC, Masaryk University, Brno

Alexiou Panagiotis *(panagiotis.alexiou@ceitec.muni.cz)*
CEITEC-MU

Baláž Andrej *(andrejbalaz001@gmail.com)*
Geneton s.r.o.

Bazgier Václav *(vaclav.bazgier@upol.cz)*
Palacký University Olomouc

Bednar Martin *(mbednar@dnanexus.com)*
DNAnexus

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

Bezděková Kateřina *(bezdekova.k@gmail.com)*
BioVendor - Laboratorní medicína a.s.

Bhagat Kriti *(501488@mail.muni.cz)*
Masaryk University

Böhmer Miroslav *(bohmerconference@gmail.com)*
Comenius University Science Park

Bojić Milan *(milan.bojic@img.cas.cz)*
CZ-OPENSCREEN, IMG

Budiš Jaroslav *(jaroslav.budis@geneton.sk)*
Geneton Ltd.

Bushuiev Roman *(roman.bushuiev@uochb.cas.cz)*
IOCB

Bystry Vojtech *(vojtech.bystry@ceitec.muni.cz)*
CEITEC MU

Čalounová Tereza *(tereza.calounova@uochb.cas.cz)*
Ústav organické chemie a biochemie AV ČR

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

Čechák David *(cechyy@gmail.com)*
CEITEC Masarykova Univerzita

Chalupová Eliška *(chalupovaeliska@email.cz)*
Masaryk University

Clarová Kamila *(kamilaclar@gmail.com)*
UCT Prague

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

Dehaen Wim *(dehaeni@vscht.cz)*
University of Chemistry and Technology, Prague

Demko Martin *(325073@mail.muni.cz)*
Masaryk university - CEITEC MU

Dohnalová Hana *(hana.dohnalova@gennet.cz)*
GENNET, s.r.o.

Ehler Edvard *(edvard.ehler@img.cas.cz)*
Ústav molekulární genetiky AV ČR, v. v. i.

Faltejsková Kateřina *(katerina.faltejskova@uochb.cas.cz)*
Ústav organické chemie a biochemie AV ČR

Feidakis Christos *(christos.feidakis@natur.cuni.cz)*
Charles University, Faculty of Science

Fišer Karel *(karel.fiser@lfmotol.cuni.cz)*
Second Faculty of Medicine, Charles University, Prague, Czech Republic

Galgonek Jakub *(jakub.galgonek@gmail.com)*
Institute of Organic Chemistry and Biochemistry of the Czech Academy of Sciences

Ghazalová Tereza *(ghazalova@enantis.com)*
Enantis s.r.o.

Giassa Ilektra-Chara *(238751@mail.muni.cz)*
CEITEC Masaryk University

Grakova Ekaterina *(ekaterina.grakova@vsb.cz)*
VSB - Technical University of Ostrava

Grešová Katarína *(514001@mail.muni.cz)*
CEITEC-MU

Halfar Radek *(radek.halfar@vsb.cz)*
VSB - Technical University of Ostrava

Hejret Václav *(vaclav.hejret@ceitec.muni.cz)*
CEITEC MU

Hloušková Lucie *(simkoval@ueb.cas.cz)*
Institute of Experimental Botany AS CR, v. v. i.

Hoksza David *(david.hoksza@matfyz.cuni.cz)*
Charles University

Holušová Kateřina *(holusovak@ueb.cas.cz)*
Institute of Experimental Botany AS CR, v. v. i.

Horáčková Jana *(jana.horackova97@gmail.com)*
Masaryk University

Janovska Anna *(veselo@volny.cz)*
CZ-OPENSCREEN IMG CAS

Jurásková Kateřina *(juraskovakaterina@seznam.cz)*
Masaryk university CEITEC

Jurič Boris *(499542@muni.cz)*
CESNET

Kiefl Yannick *(yannick.kiefl@q-ui.com)*
Ludwigs Maximilians University, Technical University of Munich, Charles University

Kolář Michal *(kolarmi@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences

Kovacova Viera *(vkovacov@uni-koeln.de)*
IBP, University of Cologne, Germany

Krampl Werner *(werner.krampl@geneton.sk)*
Geneton s.r.o.

Křenek Aleš *(ljocha@ics.muni.cz)*
CESNET

Krivak Radoslav *(rkrivak@gmail.com)*
Charles University

Kubecka Jan *(jankubecka38@gmail.com)*
Institute of Hydrobiology, Biology Center, Czech Academy of Sciences

Labala Rajendra *(michal.kolar@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences

Le Anh Vu *(lequyanh@fel.cvut.cz)*
Czech Technical University

Lepšík Martin *(lepsik@uochb.cas.cz)*
Inst. Org Chem Biochem, Czech Academy of Sciences

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

Martinek Vlastimil *(martinekvlastimil95@gmail.com)*
CEITEC MUNI

Matveieva Maria *(mariia.matveieva@upol.cz)*
VŠCHT Praha

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

Milovník Peter *(pitulep@seznam.cz)*
AbCheck s.r.o.

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

Mokrejš Martin *(mmokrejs@bioinformatics.cz)*
BIOINFORMATICS.CZ

Müller Tomáš *(tomas.muller@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences

Novotný Jiří *(fg-42@seznam.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences

Novotný Marian *(marian@natur.cuni.cz)*
Charles University

Ntie-Kang Fidele *(fidele.ntie-kang@ubuea.cm)*
University of Buea

Pačes Jan *(hpaces@img.cas.cz)*
Institute of Molecular Genetics

Pavlů Anna *(pavlua@vscht.cz)*
UCT Prague

Pešina František *(frantisek.pesina@seznam.cz)*
UCT

Petriv Oleh *(pavel.pitule@lfp.cuni.cz)*
AbCheck s.r.o.

Pfeiferová Lucie *(pfeiferoval@seznam.cz)*
IMG, Laboratory of Genomics and Bioinformatics

Pitule Pavel *(p.pitule@abcheckantibodies.com)*
AbCheck s.r.o.

Pluskal Tomáš *(tomas.pluskal@uochb.cas.cz)*
IOCB Prague

Podolská Tereza *(podolska.ter@gmail.com)*
Second Faculty of Medicine, Charles University and University Hospital Motol, Prague, Czech Republic

Popr Martin *(popr@img.cas.cz)*
Ústav molekulární genetiky AV ČR, v. v. i.

Převorovský Martin *(prevorov@natur.cuni.cz)*
Charles University, Faculty of Science

Raček Tomáš *(tom@krab1k.net)*
CEITEC Masarykova Univerzita

Říha Jakub *(jakub.riha@tul.cz)*
Technical University of Liberec

Schindler Ondřej *(ondrej.schindler@mail.muni.cz)*
Masaryk University

Schwarz Marek *(schwarz.marek@outlook.com)*
Institute of Microbiology

Sehnal David *(david.sehnal@mail.muni.cz)*
Masaryk University/CEITEC

Sieger Tomáš *(siegetom@fel.cvut.cz)*
ČVUT v Praze, Fakulta elektrotechnická

Simecek Petr *(petr.simecek@ceitec.muni.cz)*
CEITEC MU

Škuta Ctibor *(ctibor.skuta@img.cas.cz)*
IMG CAS

Sládeček Tomáš *(sladecekt@gmail.com)*
Geneton

Storchmannová Kateřina *(katerina.storchmannova01@upol.cz)*
Palacký University in Olomouc

Štourač Jan *(stourac.jan@gmail.com)*
Mezinárodní centrum klinického výzkumu Fakultní nemocnice u sv. Anny v Brně

Suchánková Pavla *(pavla.suchankova@uhkt.cz)*
Ústav Hematologie a krevní transfuze

Svatoňová Petra *(svatonovapeta@email.cz)*
Institute of Molecular Genetics

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

Svozil Daniel *(svozild@vscht.cz)*
VŠCHT Praha

Sychrovsky Vladimir *(vladimir.sychrovsky@uochb.cas.cz)*
UOCHB AV CR

Symonova Radka *(radka.symonova@gmail.com)*
Institute of Hydrobiology, Biology Center, Czech Academy of Sciences

Szemes Tomáš *(tomas.szemes@uniba.sk)*
Vedecky park Univerzity Komenskeho

Tedeschi Guglielmo *(tedeschg@vscht.cz)*
University of Chemistry and Technology

Trachtova Karolina *(k.trachtova@gmail.com)*
CEITEC

Vohradský Jiří *(vohr@biomed.cas.cz)*
Institute of Microbiology, Czech Academy of Scineces

Voršilák Milan *(milan.vorsilak@img.cas.cz)*
CZ-OPENSCREEN, IMG, CAS

Vučinić Kim *(kim.vucinic@img.cas.cz)*
Institute of Molecular Genetics of the Czech Academy of Sciences

Yurenko Yevgen *(yevgen.yurenko@gmail.com)*
IOCB Prague