XIX 3D-BIOINFO | ISCB 3D-SIG | ELIXIR CZECH REPUBLIC STRUCTURAL BIOINFORMATICS

Prague Congress Centre, November 15 - 17, 2023

Main organisers: Christine Orengo, Bohdan Schneider, Lynne Regan, Sameer Velankar, Shoshana Wodak, Vincent Zoete, Katharina Heil, Jiří Vondrášek, Anna Strachotová, Radomír Kužel

Local organization by the Czech and Slovak Crystallographic Association in collaboration with Auletris

Wednesday, November 15, Elixir Czech Republic - Session I

CL1

A PDB-WIDE ASSIGNMENT OF APO & HOLO RELATIONSHIPS BASED ON INDIVIDUAL PROTEIN-LIGAND INTERACTIONS

Marian Novotny, Christos Feidakis, Radoslav Krivak, David Hoksza

Charles University, Czech Republic
marian@natur.cuni.cz

From studying protein dynamics to unveiling cryptic binding sites or assessing the effectiveness of ligand binding site prediction software, access to several snapshots of a protein is needed. Availability of both bound (holo) and unbound (apo) forms of a protein is paramount for making meaningful comparisons and drawing robust conclusions. The few existing resources that provide access to such data are restricted either in terms of protein coverage, or in the number of provided structure pairs which does not always reflect the conformational variance that is represented by the structures deposited in the Protein Data Bank (PDB).
Here we present previously designed application (AHoJ, Apo-Holo Juxtaposition) and use it to perform an extensive search for apo-holo pairs for each individual protein-ligand interaction across the PDB (~500,000 small molecule interactions, excluding interactions with peptides and nucleic acids). We assemble the results of this search into a database that can be used to train and evaluate predictors, discover potentially druggable proteins, and reveal associations that can confirm existing hypotheses or expose protein- and ligand-specific relationships like order-to-disorder transitions, that were previously obscured by intermittent or partial data.

CL2

BINDING RESIDUE PREDICTION WITH PROTEIN LANGUAGE MODELS: DOES THE STRUCTURE MATTER?

David Hoksza, Hamza Gamouh, Marian Novotný

Charles University, Czech Republic
david.hoksza@matfyz.cuni.cz

The accurate prediction of protein-ligand binding sites is crucial for understanding protein interactions, especially in the context of biotechnology and drug discovery. There are two main approaches to tackle this challenge: one relies on the sequence of the protein (sequence-based methods), while the other relies on the three-dimensional structure of the protein (structure-based methods).

In this talk, we will discuss a novel approach that combines the strengths of both approaches to advance the state-of-the-art in this field. Our hybrid model merges two cutting-edge deep learning techniques: protein language models (pLMs) from the sequence-based approach and Graph Neural Networks (GNNs) from the structure-based approach. Specifically, we create a residue-level Graph Attention Network (GAT) model using the 3D protein structure and incorporate pre-trained pLM embeddings as node features. This integration allows our model to capture both the sequential information embedded in the protein sequence and the structural relationships within the protein.

Our model performs well compared to existing methods on a benchmark dataset, covering various ligands and ligand types. Ablation studies highlight the importance of the graph attention mechanism, particularly in densely connected graphs. Furthermore, we illustrate that as we employ more intricate pLMs to represent node features, the relative impact of the GNN architecture diminishes. This finding suggests that, to some extent, the structural information necessary for accurate binding site prediction is inherently encoded within the pLMs themselves.

CL3

NEW WAYS OF PROTEIN FAMILY VISUALIZATION IN ALPHAFOLD ERA

R. Svobodová[1,2], K. Berka[3], I. Hutařová Vařeková[1,2,3], T. Raček[1,2], O. Schindler[1,2]

[1]CEITEC – Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic
[2]National Centre for Biomolecular Research, Faculty of Science, Kamenice 5, 625 00 Brno, Czech Republic
[3]Department of Physical Chemistry, Faculty of Science, Palacký University, tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
radka.svobodova@ceitec.muni.cz

Thanks to advanced structural biology approaches, more than 200,000 experimentally determined protein structures are available in the Protein Data Bank. Based on this data, more than 200,000,000 protein structures were generated using artificial intelligence algorithms and are available in AlphaFoldDB. This data has greatly expanded the possibilities of studying protein families, their anatomy, variability, common features, and evolutionary conservation. Visualizing different aspects of protein family structures provides important information for their analysis and research.

In this lecture, we would like to present new methodologies for visualizing protein families and their properties. Specifically, 1D diagrams of protein families, generated using the OverProt tool [1], 2D diagrams, produced by the 2DProts application [2], and mapping properties and annotations of protein families onto these diagrams. The properties include, e.g., partial atomic charges calculated using the software tools ACC II [3] and Charges [4].

1. A. Midlik, I. Hutařová Vařeková, J. Hutař, A. Chareshneu, K. Berka, R. Svobodová, *Bioinf.*, **38**(14), (2022), 3648-3650.
2. I. Hutařová Vařeková, J. Hutař, A. Midlik, V. Horský, E. Hladká, R. Svobodová, K. Berka, *Bioinf.*, **37**(23), (2021), 4599-4601.
3. T. Raček, O. Schindler, D. Toušek, V. Horský, K. Berka, J. Koča, R. Svobodová, *Nucl. Acids Res.*, **48**(W1), (2020), W591-W596.
4. O. Schindler, K. Berka, A. Cantara, A. Křenek, D. Tichý, T. Raček, R. Svobodová, *Nucl. Acids Res.*, **51**(W1), (2023), W11–W16.

Core Facility Biological Data Management and Analysis of CEITEC Masaryk University, supported by ELIXIR CZ research infrastructure (MEYS Grant No: LM2023055) is gratefully acknowledged for the obtaining of the scientific data presented in this contribution.

Wednesday, November 15, Elixir Czech Republic - Session II

CL4

DYNAMICS FROM ALPHAFOLD - ELASTIC NETWORK APPROACH

Vojtech Spiwok

University of Chemistry and Technology, Prague, Czech Republic
spiwokv@vscht.cz

Alphafold 2 has significantly changed the way how structures of proteins and protein-protein complexes are being predicted. This tool also provide residue-residue distance probability profiles as its output. Using the laws of thermodynamics it is possible to infer protein dynamics from these profiles. We will present our results on application of Alphafold-based elastic network model to find flexible or conformationally variable elements in protein, for example, activation loops in protein kinases, flexible segments in G protein-coupled receptors and protein structures not known at the time of the training of Alphafold.

The work is supported by ELIXIR CZ and E-Infra (Ministry of Education, Youth and Sports of the Czech Republic: LM2018140, LM2018131) and Czech Science Foundation (22-29667S).

CL5

FIREPROT AND FIREPROT-ASR – WEB TOOLS FOR COMPUTATIONAL PROTEIN STABILISATION

David Bednar[1,2], Milos Musil[1,2,3], Rayyan Tariq Khan[1,2], Jan Stourac[1,2], Andrej Jezik[1,3], Jana Horackova[1], Simeon Borko[1,2,3], Petr Kabourek[1,2], Jiri Damborsky[1,2]

[1]Masaryk University, Czech Republic
[2]St. Anne´s University Hospital Brno, Czech Republic
[3]University of Technology, Brno, Czech Republic
davidbednar1208@gmail.com

Thermostable proteins are crucial in biomedicine and biotechnology, but designing them has been challenging, typically yielding limited improvements through single-point mutations. FireProt 2.0 builds upon its predecessor, offering several innovative strategies for protein stabilisation and starting from only a sequence input through AlphaFold integration and ProMod3 modelling. It introduces multiple-point designs with minimized antagonistic effects between mutations. Moreover, users can customize calculations, perform saturation mutagenesis for the selection of single-point mutations, or design multiple-point mutants in automized mode. Evolution-based strategies predict stabilizing mutations from back-to-consensus analysis and ancestral sequence reconstruction. FireProt 2.0 significantly reduces calculation time and improves user experience.
It is freely accessible at https://loschmidt.chemi.muni.cz/fireprot/.

Ancestral Sequence Reconstruction (ASR) infers ancestral protein sequences, aiding the discovery of highly stable, versatile, and productive proteins. FireProt^{ASR} simplifies the reconstruction process with a user-friendly web server. It automatizes the ASR process from searching homologous protein sequences, building a multiple sequence alignment, constructing and rooting the phylogenetic tree, and reconstructing the ancestral sequences, including ancestral gaps. The tool is freely available at https://loschmidt.chemi.muni.cz/fireprotasr/.

CL6

ANNOTATION, VALIDATION, REFINEMENT, AND MODELING OF NUCLEIC ACID STRUCTURES

Jiri Cerny, Paulina Bozikova, Barbora Schramlova, Bohdan Schneider

Institute of Biotechnology of the Czech Academy of Sciences, Czech Republic
jiri.cerny@ibt.cas.cz

DNATCO web server available at https://dnatco.datmos.org [1] provides intuitive annotation, validation, modeling and refinement of nucleic acids employing the universal structural alphabet of nucleic acids for assignment of DNA and RNA backbone conformations [2]. The recent improvements of the freely accessible DNATCO web server will be presented.

Further, the progress of the Nucleic Acid Valence Geometry Working Group [3], whose task is to define and implement a uniform dictionary for nucleic acid valence geometry parameters for use in modeling, refinement and validation systems will be mentioned together with initiatives dealing with standardization of nucleic acid base-pairing patterns.

1. Černý et al., *Acta Cryst*. D 2020, 76(9), 52-64, 805-813.
2. Černý et al., *Nucleic Acids Research* 2020, 48(11), 6367-6381.
3. Schneider et al., *IUCr Newsletter* 2020, 28, 4.

CL7

ARE KURAVIRUS CAPSID DIAMETERS QUANTIZED? THE FIRST ALL-ATOM GENOME TRACING METHOD FOR DOUBLE-STRANDED DNA VIRUSES

Samuel Coulbourn Flores[1], Michal Malý[2], Dominik Hrebík[3], Pavel Plevka[3], Jiří Černý[2]

[1]Swedish University of Agricultural Sciences, Sweden
[2]Institute of Biotechnology of the Czech Academy of Sciences
[3]Central European Institute of Technology, Brno, Czech Republic
samuel.flores@slu.se

The revolution in Cryo-Electron Microscopy has resulted in unprecedented power to resolve large macromolecular complexes including viruses. Many methods exist to explain density corresponding to proteins and thus entire protein capsids have been solved at the all-atom level. However methods for nucleic acids lag behind, and no all-atom viral double-stranded DNA genomes have been published at all. We here present a method which exploits the spiral winding patterns of DNA in icosahedral capsids. The method quickly generates shells of DNA wound in user-specified, idealized spherical or cylindrical spirals. For transition regions, the method allows guided semi-flexible fitting. For the *kuravirus* SU10, our method explains most of the density in a semiautomated fashion. The results suggest rules for DNA turns in the end caps under which two discrete parameters determine the capsid inner diameter. We suggest that other kuraviruses viruses may follow the same winding scheme, producing a discrete rather than continuous spectrum of capsid inner diameters. Our software may be used to explain the published density maps of other double-stranded DNA viruses and uncover their genome packaging principles.