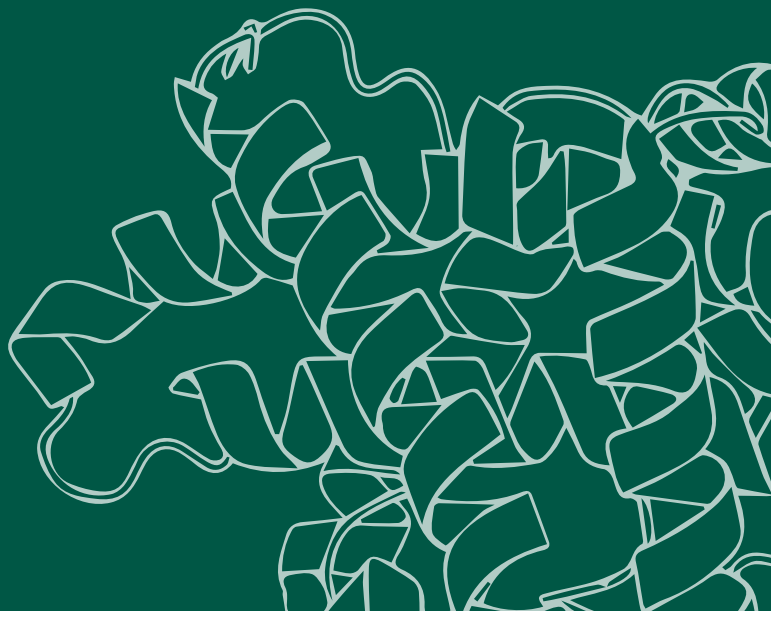
EUROPEAN ROSETTACON
ON PROTEIN STRUCTURE
PREDICTION AND DESIGN

WARSAW
10-13
MAY
2022

BOOK OF ABSTRACTS



Rosetta Commons

UNIVERSITY OF WARSAW

2022 European Rosetta Conference

10 May-14 May 2022

Keynote Abstract

Protein design using deep learning

David Baker
University of Washington, IPD

Proteins mediate the critical processes of life and beautifully solve the challenges faced during the evolution of modern organisms. Our goal is to design a new generation of proteins that address current-day problems not faced during evolution. In contrast to traditional protein engineering efforts, which have focused on modifying naturally occurring proteins, we design new proteins from scratch to optimally solve the problem at hand. We now use two approaches. First, guided by Anfinsen's principle that proteins fold to their global free energy minimum, we use the physically based Rosetta method to compute sequences for which the desired target structure has the lowest energy. Second, we use deep learning methods to design sequences predicted to fold to the desired structures. In both cases, following the computation of amino acid sequences predicted to fold into proteins with new structures and functions, we produce synthetic genes encoding these sequences, and characterize them experimentally. In this talk, I will describe recent advances in protein design using both approaches.

Talk Abstracts

Making a completely de-novo designed random protein walker/roller

Ajasja Ljubetič, Joseph Watson, Hao Shen, Natasha I Edman, David Baker
National Institute of Chemistry, Department for Synthetic Biology and Immunology

Powered protein walkers such as kinesin, dynein or myosin are responsible for most movements within the cell and for the transport of crucial cargo. Design of static monomeric and oligomeric protein structures has advanced tremendously; however large dynamic protein mechanisms have not yet been designed. I will present the design and characterization of a random protein walker that can diffuse along micrometer long fibers. This represents a scaffold for future powered molecular robots. The requirements for such a system are threefold: a track, attachment points and a walker/roller scaffold. I will briefly present reversible heterodimers that I developed to serve as attachment points. These heterodimers behave well as monomers, have a range of affinities and fast binding/exchange kinetics in solution. Next, I will show how I rigidly fused heterodimers and designed helical repeat (DHR) proteins to de-novo designed fibers to form a track for the walker. This turned out to be the hardest part of the project. Finally, I will outline the different walker/roller scaffolds I used and show trajectories obtained from single molecule microscopy experiments.

Expanding the repertoire of de novo protein assemblies: secretion optimized and polymorphic assemblies

Alena Khmelinskaia, John Wang, Neil P. King
Life and Medical Sciences Institute, University of Bonn,Gerhard-Domagk-Str. 1

Protein assemblies have long been targeted by the nanomaterial community for their various functions in nature, such as shielding macromolecules from the surrounding environment and providing spatial control over biochemical reactions. Recently, computational methods have been developed for designing novel protein assemblies with atomic-level accuracy, yet several aspects of current methods limit the structural and functional space that can be explored. For example, the designed hydrophobic interfaces that are essential for successful assembly are often interpreted by cells as transmembrane segments, resulting in inefficient secretion. Furthermore, the underlying perfect symmetry limits the size and types of architectures that can be designed. Here, I will discuss approaches we have developed to overcome these limitations. First, I will describe the "degreaser", a new computational protocol developed to not only improve secretion of natural proteins and existing protein materials, but also streamline the design of novel SOAPs (secretion optimized assembling proteins). Next, I will discuss how local instabilities introduce structural flexibility in protein scaffolds, breaking symmetry and vastly expanding the repertoire of possible target architectures.

Matching protein surface structural patches for peptide docking

Alisa Khramushin

The Hebrew University of Jerusalem, Department of Microbiology and Molecular Genetics, Institute for Biomedical Research Israel-Canada, Faculty of Medicine

Modeling interactions between short peptides and their receptors is a challenging docking problem due to the peptide flexibility, resulting in a formidable sampling problem of peptide conformation in addition to its orientation. Alternatively, the peptide can be viewed as a piece that complements the receptor monomer structure. We show that the peptide conformation can be determined based on the receptor backbone only and sampled using local structural motifs found in solved protein monomers and interfaces, independent of sequence similarity. This approach outperforms current peptide docking protocols and promotes new directions for peptide interface design.

Induced fit with replica exchange improves protein complex structure prediction

Ameya Harmalkar
Johns Hopkins University, Baltimore, MD

Despite the progress in prediction of protein complexes over the last decade, recent blind protein complex structure prediction challenges revealed limited success rates (less than 20% models with DockQ score >0.4) on targets that exhibit significant conformational change upon binding. To overcome limitations in capturing backbone motions, we developed a new, aggressive sampling method that incorporates temperature replica exchange Monte Carlo (T-REMC) and conformational sampling techniques within docking protocols in Rosetta. Our method, ReplicaDock 2.0 , mimics induced-fit mechanism of protein binding to sample backbone motions across putative interface residues on-the-fly, thereby recapitulating binding-partner induced conformational changes. Furthermore, ReplicaDock 2.0 clocks in at 150-500 CPU hours per target (protein-size dependent); a runtime that is significantly faster than Molecular Dynamics based approaches. For a benchmark set of 88 proteins with moderate to high flexibility (unbound-to-bound iRMSD over 1.2 A), ReplicaDock 2.0 successfully docks 61% of moderately flexible complexes and 35% of highly flexible complexes. Additionally, we demonstrate that by biasing backbone sampling particularly towards residues comprising flexible loops or hinge domains, highly flexible targets can be predicted to under 2 A accuracy. This indicates that additional gains are possible when mobile protein segments are known.

Organic Anion Transporter 1: Insights into function and structure by means of Molecular Dynamic simulations

Angelika Janaszkiewicz, Ágota Tóth, Quentin Faucher, Marving Martin, Benjamin Chantemargue, Florent Di Meo

INSERM U1248 Pharmacology & Transplantation, University of Limoges

The human Organic Anion Transporter 1 (hOAT1) is known to play a central role in renal drug elimination and the maintenance of systemic homeostasis. Since protein-mediated drug transport across cell membrane affects local pharmacokinetics (PK), the structural understanding of transporter function has become of emerging importance. Despite its significance, the knowledge about the structure and the transport mechanism of hOAT1 remains fragmented, owing to the lack of resolved structure. The present study provides the first robust model of hOAT1 in outward-facing conformation obtained by protein-threading techniques regined by μs-scaled Molecular Dynamics (MD) simulations. Such approach allowed to reconstruct the well-known structural arrangement of transmembrane helices observed in Major Facilitator Superfamily. Besides, the inward-facing conformation of hOAT1 was built from the AlphaFold 2 structure prediction tools which was also used for further MD simulations. By combining simulations of both OF and IF conformations, essential structural and functional features are provided. Embedding hOAT1 models into different lipid bilayer membranes revealed insights into protein-lipid interactions in which phosphatidylethanolamine (PE) components were suggested to play an active role in hOAT1 function. For instance, protein-lipid interactions were shown to participate to the allosteric communication between intra- and extracellular domains upon substrate binding events. Furthermore, hOAT1 shares intracellular motifs which are conserved among MFS proteins. They were found to form a complex pseudo-symmetrical network of salt-bridges, likely to be crucial in the large-scale conformational change along transport cycle. The present MD-refined models were also used to provide hints about pharmacologically relevant defects arising from single nucleotide polymorphism.

Molecular Dynamics-based descriptors of 3-O-Sulfated Heparan Sulfate as Contributors of Protein Binding Specificity

Annemarie Danielsson, Malgorzata M. Kogut, Martyna Maszota-Zieleniak, Pradeep Chopra, Geert-Jan Boons, Sergey A. Samsonov

University of Gdansk, Department of Theoretical Chemistry, Faculty of Chemistry

Glycosaminoglycans (GAGs) are linear periodic and anionic polysaccharides found in the extracellular matrix, involved in key biochemical processes as a result of their interactions with a variety of protein partners. Due to the template-less synthesis, high flexibility and charge of GAGs, as well as the multipose binding of GAG ligands to proteins, the specificity of GAG-protein interactions can be difficult to elucidate. In our study, we analyzed the specificity of Heparan Sulfate (HS)-protein interactions employing computational approaches - we applied molecular dynamics (MD) simulations as well as a simple and well-known machine learning algorithm (Principal Component Analysis) to examine HS hexasaccharides of different sulfation patterns, followed by linear regression and cluster analysis in order to assess the connection between characteristics of unbound HS molecules and the specificity of HS-protein binding. The investigation of the behavior of unbound HS molecules during MD simulations at atomistic level, supported and complemented by experimental data on GAG-protein binding affinity, allowed us to identify characteristics of unbound HS molecules that could be linked with confidence to differences in binding affinity for a set of HS-protein complexes.

De novo design of protein interactions with learned surface fingerprints

Anthony Marchand, Pablo Gainza, Sarah Wehrle, Alexandra K. Van Hall-Beauvais, Andreas Scheck and Bruno E. Correia

Ecole polytechnique fédérale de Lausanne, Laboratory of protein design and immunoengineering

Protein-protein interactions (PPI) play a major role in various biological processes in healthy cell homeostasis and disease progression. Due to the complex interplay of energetic contributions that drive specific molecular recognition events, the design of de novo protein interactions remains challenging. Here, we proposed a new computational method based on learned surface fingerprints for the generation of de novo PPIs against various targets of interest. Recently, our group developed a geometric deep-learning framework, MaSIF (Molecular Surface Interaction Fingerprinting), for predicting PPI interfaces and partners based on vectorized geometric and chemical features of the protein surface, also referred to as fingerprints. Our trained algorithm showed reliable predictions of existing PPIs three to four orders of magnitude faster than other state-of-the-art algorithms. Our group has since enhanced MaSIF for the design of de novo PPIs based on surface fingerprint matching. In this work, we used MaSIF to search for binding seeds on which an interface can be further designed. Binding seeds were identified from a large helical protein fragment database and grafted onto scaffold proteins that were further computationally optimized with Rosetta to improve global properties favorable for binding interactions. Selected designs were then screened using high-throughput approaches, improved by experimental maturation techniques if necessary and characterized with various biophysical methods for folding, stability and binding affinities. As a proof–of-principle we successfully designed and tested four de novo protein binders to engage three protein targets: SARS-CoV-2 spike, PD1, and PDL1. Altogether, this work will serve as a basis for an innovative PPI design strategy based on surface fingerprinting and a hotspot-centric approach. Ultimately, novel PPIs produced by this work will contribute to a better understanding of PPI design and open the possibility of developing innovative biologics or cell-based therapies.

Reading and Writing Protein Function Using Multi-dimensional Surface Representations

Bruno E Correia

Ecole Polytechnique Federale de Lausanne, Institute of Bioengineering

Predicting interactions between proteins and other biomolecules solely based on structure remains a central challenge in biology. A high-level representation of protein structure, the molecular surface, displays patterns of chemical and geometric features that fingerprint a protein's modes of interactions with other biomolecules. The underlying hypothesizes of our work is that proteins interacting with similar molecules may share common fingerprints, independent of their sequence and overall structural fold. These structural fingerprints are difficult to grasp by visual analysis but may be learned from large-scale datasets.

I will discuss MaSIF (Molecular Surface Interaction Fingerprinting), a conceptual framework based on a geometric deep learning method, to capture structural fingerprints that are important for specific biomolecular interactions. I will also present current developments of the framework to make it end-to-end differentiable. Finally, I will describe our approach based on MaSIF to design de novo protein-protein interactions, which may have important applications in the development of new protein-based drugs.

We anticipate that this conceptual framework will lead to improvements in our understanding of protein function and design.

De novo design of dynamic proteins

Florian Praetorius, Phil Leung, and David Baker
Baker Lab, Institute for Protein Design,University of Washington, Seattle

Computational design of a novel protein normally focuses on identifying the optimal amino acid sequence for a single stable conformation. The resulting proteins are usually very stable across a broad range of temperatures, solution conditions, and even sequence variations. While this stability is a useful feature for many applications of designed proteins, the structural rigidity of these proteins can also be limiting. Many natural proteins rely on conformational dynamics to perform their functionalities. Most notably, molecular motors couple chemical reactions with conformational changes, enabling the transformation of chemical energy into mechanical energy. Here, we are designing proteins that can adopt two or more well-defined conformational states. Starting from previously designed proteins with experimentally verified structures we identify potential alternative conformations by sampling rigid-body movement of individual domains. For selected alternative conformations we then use a multi-state design approach to identify sequences that are compatible with both the alternative state and the original conformation of the input protein. *Ab initio* structure prediction can then be used to assess the fitness of the resulting sequence for the designed conformations, and to check for potential off-target states. We are also designing proteins in which the alternative state of the protein can bind to a target molecule, such as a helical peptide, a small molecule, or another copy of itself. In these systems, binding and conformational change are coupled. We envision that these proteins can then be used to build reconfigurable protein complexes, logic devices, or even molecular motors. Here, we are designing proteins that can adopt two or more well-defined conformational states. Starting from previously designed proteins with experimentally verified structures we identify potential alternative conformations by sampling rigid-body movement of individual domains. For selected alternative conformations we then use a multi-state design approach to identify sequences that are compatible with both the alternative state and the original conformation of the input protein. *Ab initio* structure prediction can then be used to assess the fitness of the resulting sequence for the designed conformations, and to check for potential off-target states.

Disclosing the binding properties of pet tracers targeting tau by computational simulations

Georg Künze

Universität Leipzig, Medizinische Fakultät, Institut für Wirkstoffentwicklung

The abnormal deposition of hyperphosphorylated tau protein into intraneuronal neurofibrillary tangles in the human brain is a pathological hallmark of Alzheimer's disease (AD). Tau deposits are also causative for other neurodegenerative diseases such as progressive supranuclear palsy (PSP), and corticobasal degeneration (CBD). AD is the most common cause of dementia, accounting for an estimated 60-80% of cases in the world. Noninvasive imaging of tau fibrils by positron emission tomography (PET) can improve the early diagnosis of AD and other tau-related pathologies by detecting early changes in pathological tau levels in cognitively unimpaired individuals. [18F]PI-2620 is a novel tau-PET tracer that detects tau fibrils in AD, CBD, and PSP with distinct binding characteristics. Higher clearance of [18F]PI-2620 in cases of CBD and PSP indicates less stable tracer binding in these tauopathies compared to AD. To obtain mechanistic insight into the interaction mode of [18F]PI-2620 with tau fibrils and the structural basis for the distinct binding characteristics of AD, CBD, and PSP tau fibrils, a multiscale simulation workflow was applied. Molecular docking and metadynamics simulations were used to exhaustively search for possible tracer binding sites and extensively sample the free energy landscape of tracer binding. Microsecond molecular dynamics simulations and MM/GBSA energy calculations provided information about tracer flexibility and the binding free energy of each tracer binding site. Finally, Brownian Dynamics simulations were carried out to determine the association kinetics of the tracer-tau fibril complexes. While in AD-tau fibrils the most favorable binding sites of [18F]PI-2620 are located on the concave side of the fibril surface, the lowest-energy binding sites in CBD and PSP-tau fibrils are found at the inner cavity of the fibril core structure. Tracer association rates for surface sites are found to be higher than association rates at cavity sites. These results suggest that tracer binding sites in AD fibrils have high loading capacity due to favored energetic and kinetic properties, whereas intercalation of tracer molecules into the inner cores of CBD and PSP fibrils is kinetically hindered. Detailed structural knowledge like this will be critical to develop strategies in tau-PET imaging that can differentiate between AD, CBD, and PSP, and design novel tau-PET tracers with high selectivity for certain tauopathies.

Antibody Engineering with Deep Learning

Jeff Gray, Jeff Ruffolo, Richard Shuai, Jeremias Sulam, Sai Pooja Mahajan, Lee-Shin Chu

Johns Hopkins University, Chemical & Biomolecular Engineering

In addition to their role in the immune system, antibodies are important therapeutic molecules. In this talk I will summarize work in the lab using deep learning approaches for antibody engineering. I will discuss our fast and accurate language models for antibody structure prediction, a generative model for creation of antibody libraries with favorable developability characteristics, and hallucination approaches for antibody-antigen interface design.

Engineering antibodies and vaccines with Rosetta

Jens Meiler

Leipzig University, Medical Faculty,Institute for Drug Discovery

Engineering the optimal human antibodies to fight viral infections is important in order to guide epitope-focused vaccine design. On several example viruses including HIV, Influenza, Marburg, Ebola, and Corona a number of innovative recent computational technologies developed for Rosetta will be introduced. Specifically, computationally designed cyclic peptides derived from an antibody loops will be introduced, it will be shown that multi-state design of viral proteins predicts sequences optimal for conformational change, the discovery of neutralizing antibodies from virus-naive human antibody repertoires using large-scale structural prediction will be discussed, and an algorithm for the design of protein multi-specificity using an independent sequence search that reduces the barrier to low energy sequences will be analyzed.

Can we determine which enzyme designs will be catalytic without experimentation?

Ryan Feehan, Meghan Franklin, Joanna Slusky

University of Kansas, Molecular Biosciences and Computational Biology

I will present the recent development of a machine learning model to distinguish between catalytic and non-catalytic metalloprotein sites.

Using Rosetta to find small molecules that rescue destabilized protein mutants

John Karanicolas, Sven Miller, Chris Parry, Mariam Fouad Hafez
Fox Chase Cancer Center, Molecular Therapeutics Program

Clinical genetics points to many human diseases for which the underlying pathology can be traced to mutations that map to the interior of a folded protein: we hypothesize that these mutations act by destabilizing an otherwise folded protein, such that the protein loses activity because an insufficient amount of the cellular population is correctly folded. However, current drug discovery expertise is centered around inhibitors of enzyme activity and modulators of cell surface receptors, leaving the field ill-equipped to tackle the challenge of designing compounds that restore the function of proteins deactivated in this manner. The Karanicolas group has been developing tools in Rosetta specifically catered to the shallow surface pockets typical of sites that are not naturally evolved for small-molecule binding. We have recently applied these tools to identify novel druggable sites on the surfaces of two different tumor suppressor proteins, and using Rosetta for virtual screening we identified small drug-like molecules that bind to these sites. By treating cancer cell lines that harbor these mutant tumor suppressors with the corresponding stabilizers, we find that these compounds refold these destabilized mutant proteins and restore WT activity. We anticipate that the compounds described here may serve as a starting point for new classes of cancer therapeutics. More broadly, however, these represent first proof-of-concept for a new therapeutic modality: using small molecules to revert loss-of-function induced by mutations that act by disrupting protein stability.

More than Proteins: Programming Proteins Using Non-Protein Components

Jonathan G. Heddle

Malopolska Centre of Biotechnology, Jagiellonian University

Protein design tools have allowed the predictive design and production of diverse, often functional structures. Nature itself demonstrates an impressive range of capabilities achievable by protein machines. However, in the vast space of possible functionalities it may be that there are some which will remain impossible or at least impracticable to achieve using protein alone. In such cases we can include non-protein components into designed protein structures to provide additional functionality. In this presentation I will describe our own tentative first steps in this area where we have succeeded in incorporating non-protein molecules into an artificial cage structure such that the conditions in which the cage breaks apart can be controlled.

Hydrogen Bond Potential in course-grained force field

Justyna Kryś
Faculty of Chemistry, UW

Hydrogen bonding is one of the most important interactions that are responsible for forming native three-dimensional protein structures. They are also the driving force that creates secondary structure elements as helices and sheets, which make proteins much different from random chains of polymers. Hydrogen bonds therefore has played a crucial role in force-fields used for protein simulations.
A hydrogen bond is formed when a proton covalently attached to one electronegative donor atom is shared with another electronegative acceptor atom. This interaction in general has a semi-covalent character that should be described at quantum chemistry level. Numerous all-atom approaches have been proposed to provide an accurate mathematical formulation describing this interaction as a function of distances as well as planar and dihedral angles measured between the atoms involved.
In coarse-grained modeling the position of hydrogen donors and acceptors are often not explicitly defined which makes such a description even more difficult. During my talk I will present novel approach of hydrogen bonds term that can be used even if only CA trace is available. By the careful statistical analysis of know protein structures it was possible to derive a potential of mean force that can correctly identify hydrogen bonds and asses their geometry.

Fungal genome mining for a super cytochrome P450 enzyme for higher higher-molecular-weight polycyclic aromatic hydrocarbons degradation

Khajamohiddin Syed

Department of Biochemistry and Microbiology, Faculty of Science and Agriculture, University of Zululand

Polycyclic aromatic hydrocarbons (PAHs) and long-chain alkylphenols (Aps) are known mutagens/carcinogens. Cytochrome P450 monooxygenases (CYPs/P450s) play a key role in detoxifying/removing these pollutants. A study revealed the presence of a P450 enzyme, CYP63A2, from the model white-rot fungus *Phanerochaete chrysosporium*, with catalytic versatility of oxidation of up to six-ring PAHs and a range of alkylphenols. Structural analysis revealed that CYP63A2 has the largest active site compared to humans (CYP3A4, CYP1A2, and CYP1B1) and bacterial (CYP101D and CYP102A1), indicating it can accommodate larger molecules. Genome-wide data mining and annotation of CYP63 members in fungi were performed to find a super P450 with magnificent catalytic versatility. This study revealed the presence of CYP63 members in most of the fungal species. Based on the evolutionary pattern of CYP63 members, potential CYP63 proteins were selected. Virtual screening of these potential CYP63 members' ; substrate affinities against PAHs and Aps, is in progress.

Cyclic peptide approach to model protein-glycosaminoglycan interactions

Margrethe Gaardløs[1], Brianda L. Santini[2], Martin Zacharias[2], and Sergey A Samsonov[1]

Glycosaminoglycans (GAGs) are ubiquitous anionic periodic linear polysaccharides in the extracellular matrix in animal tissue. Protein-GAG interactions regulate many fundamental processes, including cell growth and differentitation, anticoagulation, inflammation and cancer progression. Due to their extensive physiological effects, many of these interactions are important therapeutical targets in medicine and drug design. One example is the interaction between the serin protease inhibitor antithrombin and the glycosaminoglycan heparin. Antithrombin itself is used therapeutically as an anticoagulant, and its action is upregulated from the interaction with a specific heparin pentasaccharide sequence. The synthetic anticoagulant Fondaparinux is based on this sequence. Rational design of therapeutics that mimic binding segments requires an understanding of the molecular mechanisms involved. However, the interactions are difficult to study computationally, due to properties including the flexibility and periodicity of GAGs, their highly charged nature, and the lack of binding pockets on the protein surfaces where they bind. New in silico approaches that could aid in characterization of these interactions and the design of mimics are therefore valuable. Previously, a novel theoretical method was developed to rationally design cyclic peptides mimicking the binding site of one of the interaction partners in protein-protein interactions (PPIs) [1]. This approach automatically characterizes a protein binding site in terms of backbone coordinates, and finds a matching cyclic peptide by searching through a library of cyclic peptide backbone structures. Compared to linear peptides, cyclic peptides form rigid structures that are less susceptible to proteolysis. For the first time, this approach was used for protein-GAG interactions, specifically for the complex of antithrombin and a heparin pentasaccharide. Just like for PPIs, the protein partner in these interactions have relatively flat binding surfaces without defined pockets. The resulting cyclic peptide-GAG complexes were investigated through conventional and enhanced sampling molecular dynamics simulations in which stable complexes were observed with several different cyclic peptides. The approach involves an automated characterization of the GAG binding sites, which is used for a thorough investigation of motifs in binding sites of known GAG-binding proteins. A modified version of this approach could also be used to find cyclic peptides that mimic GAGs, with further potential applications for drug design.

[1] Santini BL, Zacharias M. Rapid in silico Design of Potential Cyclic Peptide Binders Targeting Protein-Protein Interfaces. Frontiers in Chemistry. 2020;8:933.

Multi-Scale Flexible Fitting of Proteins to Cryo-EM Density Maps

Marta Kulik [1,2] , Takaharu Mori [2], Yuji Sugita [2]

[1] Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw, Poland [2] RIKEN Cluster for Pioneering Research, Saitama, Japan

Despite the tremendous progress in protein structure prediction in silico, Cryo-EM at medium resolution is still instrumental in the structure determination of biological complexes of proteins. To obtain a full-atom protein model based on the Cryo-EM density maps, flexible fitting molecular dynamics simulations have proven to be useful. Here, we present a new protocol for flexible fitting of proteins to Cryo-EM density maps [Kulik M, Mori T and Sugita Y, Front. Mol. Biosci. 2021], able to avoid overfitting, automatically adjust the force constants driving the structure to the density map, and correctly recreate the complex conformational transitions of protein domains. The protocol consists of three steps. First, we perform coarse-grained flexible fitting molecular dynamics simulations with a replica-exchange scheme between different force constants. Then, in targeted molecular dynamics simulation, the fitted Cα atom positions from the first step guide the all-atom structure to the correct positions. Finally, the structure is refined with the all-atom flexible fitting simulation in implicit solvent. The last step corrects the side-chain arrangement. All in all, the final models obtained via the multi-scale protocol are well resolved, reliable, and do not contain overfitted regions in comparison with the simple all-atom flexible fitting simulations.

This work was supported by the RIKEN Pioneering Project Dynamic Structural Biology and MEXT/Kakenhi (Grant number 19H05645 to YS). MD simulations were carried out at RIKEN on HOKUSAI BigWaterFall.

Protein Modeling and Design for Enzymatic Conversion of Universal Blood

Morgan L. Nance, Peter Rahfeld, Charlotte Olagnon, Stephen G. Withers, Jeffrey J. Gray

Johns Hopkins University, Program in Molecular Biophysics

Whole blood (WB) transfusion is an indispensable pre-hospital and perioperative intervention to improve survival outcomes of patients suffering severe hemorrhage. A critical safety consideration is the donor-recipient blood type compatibility given that a typing mismatch can result in hemolysis and, in severe cases, patient death. ABO blood type is determined in part by the presence or absence of specific, recognizable carbohydrate antigens on the cell surface of red blood cells (RBCs); Group A blood has A-antigens, Group B has B-antigens, Group AB has both, and Group O has neither. Accordingly, Group O WB is the only choice for hemostatic resuscitation in emergency situations when the blood type of the patient is unknown. Despite Group O being the most common blood type in the U.S., there is a constant need for Group O WB due to only 5% of eligible donors actually donating blood. Consequently, the enzymatic conversion of Group A and B RBCs to O-type (Enzyme-Converted Group O RBCs or ECO-RBCs) serves as a historically feasible and potentially practical solution to meet the constant demand for "universal" Group O WB for lifesaving patient care. Here, I will discuss how the Rosetta macromolecular modeling and design software suite is used to model the full, multi-domain structures of a pair of enzymes capable of generating ECO-RBCs from Group A WB. I will also describe how GlycanDock (my recently developed protein-glycoligand docking refinement algorithm) is used to predict the bound conformation of different enzyme-carbohydrate blood antigen complexes. My structural models inform further Rosetta simulations to find mutations that increase the stability and expression of the two enzymes. Finally, I will describe my protein design efforts to re-engineer one of the enzymes to enable Group B WB conversion.

A Deep unsupervised Language Model for Protein Design

Noelia Ferruz, Steffen Schmidt, Birte Höcker
University of Girona, University of Bayreuth

Protein design aims to build new proteins from scratch thereby holding the potential to tackle many environmental and biomedical problems. Recent progress in the field of natural language processing (NLP) has enabled the implementation of ever-growing language models capable of understanding and generating text with human-like capabilities. Given the many similarities between human languages and protein sequences, the use of NLP models offers itself for predictive tasks in protein research. Motivated by the evident success of generative Transformer-based language models such as the GPT-x series, we developed ProtGPT2, a language model trained on protein space that generates de novo protein sequences that follow the principles of natural ones. In particular, the generated proteins display amino acid propensities which resemble natural proteins. Disorder and secondary structure prediction indicate that 88% of ProtGPT2-generated proteins are globular, in line with natural sequences. Sensitive sequence searches in protein databases show that ProtGPT2 sequences are distantly related to natural ones, and similarity networks further demonstrate that ProtGPT2 is sampling unexplored regions of protein space. AlphaFold prediction of ProtGPT2-sequences yielded well-folded non-idealized structures with embodiments as well as large loops and revealed new topologies not captured in current structure databases. ProtGPT2 has learned to speak the protein language. It has the potential to generate de novo proteins in a high throughput fashion in a matter of seconds. The model is easy-to-use and freely available.

Peptide binding as monomer complementation - new approaches for blind global high-resolution peptide docking

Ora Schueler-Furman

Hebrew University of Jerusalem, Faculty of Medicine

Peptide-mediated interactions play crucial roles in cellular regulation. Challenged by the flexibility of the peptide on the one hand and the often transient and weak character of the interaction on the other, they pose special challenges, both for modeling and experimental efforts. Recent advances in Deep Learning, such as by Deepmind Alphafold2, are revolutionizing computational structural biology, bringing to reach high accuracy models of full proteomes. Evidence has accumulated that the binding of peptides to their receptors could be seen as monomer complementation, i.e., a final step of monomer folding. Based on this concept, we have developed two novel, top-performing peptide docking protocols as assessed on a comprehensive benchmark and validated set. The first, PatchMAN(1), applies a fast search to match patches on the receptor surface for structural motifs in solved structures. These can then be used to extract complementing fragments that seve as starting point for peptide refinement. The second approach uses a slight modification of Alphafold2 to model the peptide either as separate unit, or connected by a poly-glycine linker to the c-terminus of the receptor(2). Importantly, we succeed in modeling these interactions at high accuracy, even though no information on the multiple sequence alignment of the peptide partner is available.

I will introduce these approaches and their performance, and then discuss the underlying reasons for success, and failure. This can teach us about the basic principles of this interesting and important type of interactions between proteins.

(1) Alisa Khramushin, Ziv Ben-Aharon, Tomer Tsaban, Julia K Varga, Orly Avraham, Ora Schueler-Furman (2022). Matching protein surface structural patches for high-resolution blind peptide docking. Accepted for publication in PNAS. Original version available on Biorxiv doi: 10.1101/2021.09.02.458699

(2) Tomer Tsaban, Julia Varga, Orly Avraham Ziv Ben-Aharon, Alisa Khramushin, Ora Schueler-Furman (2022) Harnessing protein folding neural networks for peptide–protein docking. Nat Commun 2022, 13:2021.08.01.454656.

Understanding the peptide folding landscape using computational methods

Parisa Hosseinzadeh, Noora Azadvari, Suchetana Gupta
University of Oregon,1505 Franklin Boulevard, M343

Despite recent success of deep learning methods in predicting protein structures, predicting structure of peptides in solution is still a challenging task. This challenge is in part due to the unnatural features of peptides such as the use of crosslinkers and non-canonical building blocks which limits the generalizability of protein structure prediction methods to peptides. Another challenge is the fact that peptides often can take on multiple conformations in solution and thus methods that are focused on finding one global minimum in the folding landscape of peptides will not be able to capture the true behavior of peptides in solution. To address these challenges, we use a combination of Rosetta based conformational sampling, molecular dynamics simulations, and machine learning to better understand the folding landscape of peptides and therefore, better predict their behavior in solution.

Computational design of signaling membrane receptors

Patrick Barth

EPFL, Institute of Bioengineering

The ability to dynamically switch between distinct conformations and transduce long-range signals represents a hallmark of membrane receptor functions. Understanding the molecular underpinnings of these critical activities remains however challenging as subtle differences in protein sequence often give rise to profound changes in signaling response with no obvious connection to their structure. We developed a computational approach for predicting and designing allosteric signaling functions. Using the method, we designed G protein-coupled receptors (GPCRs) with novel ligand binding and signaling selectivity that agreed well with our predictions. Combining allostery and de novo design techniques, we created stabilized functional variants of challenging GPCRs and characterized their structure in active signaling states. We generalized the approach and created potent single- and multipass membrane receptor biosensors that reprogrammed immune cell functions. Our work should prove useful for engineering a wide range of biosensors with programmable sensing-signaling behaviors for basic and therapeutic applications.

Epitope-specific antibody design by a deep learning generative model

Raphael R. Eguchi, Christian A. Choe, Possu Huang
Stanford University, Shriram Center for Chemical Engineering and Bioengineering

The growing need for antibodies with customized specificity provides a rich environment for engineering efforts. In recent years, despite having streamlined experimental pipelines, the fundamental math requiring extensive libraries and screen campaigns to get an initial binding signal remains unchanged. A major advancement would be to directly design *in silico* an epitope-specific binder from scratch, providing a signal for potential optimization by artificial evolution. We have observed several key advantages in neural network approaches over existing methods. By leveraging the unique properties of neural networks, we developed a generative model for immunoglobulin 3D structures, with which diverse structures can be modeled with unprecedented speed. We extended it to a purely deep learning-based protein-protein interface design pipeline that optimize not only spatial orientations but fully-flexible protein structures on the fly to desired epitopes. This novel strategy explores neural network's capabilities in modeling dynamic structures, and preliminary experimental results on multiple targets support the plausibility of *in silico* design of epitope-specific antibodies.

Computational Protein Structure Prediction from Mass Spectrometry Data

Steffen Lindert

Ohio State University, Department of Chemistry and Biochemistry

Mass spectrometry-based methods such as covalent labeling, surface induced dissociation (SID) or ion mobility (IM) are increasingly used to obtain information about protein structure. However, in contrast to other high-resolution structure determination methods, this information is not sufficient to deduce all atom coordinates and can only inform on certain elements of structure, such as solvent exposure of individual residues, properties of protein-protein interfaces or protein shape. Computational methods are needed to predict high-resolution protein structures from the mass spectrometry (MS) data. Our group develops algorithms within the Rosetta software package that use mass spectrometry data to guide protein structure prediction. These algorithms can incorporate several different types of mass spectrometry data, such as covalent labeling, surface induced dissociation, and ion mobility. We developed scoring functions that assess the agreement of residue exposure with covalent labeling data, the agreement of protein-protein interface energies with SID data and the agreement of protein model shapes with collision cross section (CCS) IM measurements. We subsequently rescored Rosetta models generated with de novo protein folding and protein-protein docking and we were able to accurately predict protein structure from MS labeling, SID and IM data.

Designing stable metalloproteins using deep learning without force field

Simon L. Dürr, Andrea Levy, Ursula Röthlisberger
EPFL SB ISIC LCBC, BCH 4109

30 to 40 % of proteins are estimated to depend on at least one metal ion for their biological function1. Despite their important biological function the computational design of metalloproteins remains an arduous task due to the inaccuracy of force fields for metals (especially for important transition metals such as zinc) and the computational cost of QM calculations for biologically relevant systems. Some computational successes include the design of a zinc-mediated PPI which due to the inaccuracy of the scoring function exhibited a vacant site and a small cavity close to it and had low esterase activity[2] or our redesign of GB1 which introduced a metal site with one vacancy but dimerized in solution as head-to-tail dimer via the Zn2+ binding site[3]. While designing simple binding sites rationally is something that has been achieved for multiple different folds, computational design of functional metallosites with defined first and second shell coordination such as required for enzymatic activity has not been achieved yet[4]. In this work, we present applications of deep learning towards designing stable metalloproteins using 3D convolutional neural networks. Similar networks have been used for predicting binding pockets in proteins[5], for identification of masked residues[6] and for fixed backbone protein design[7]. In this work, we present new results for the prediction of masked residues with metals as an explicit input channel and evaluate the performance with respect to preorganized active sites of metalloenzymes. A comparison with deep mutational scanning results, $\Delta\Delta$G predictions and Zn2+ binding assays shows that the network can filter out false-positive $\Delta\Delta$G predictions efficiently and captures important hydrogen bond networks inside the active site of natural metalloenzymes and thus might be useful to optimize the second shell around metal active sites. Secondly, we present a network that, given a protein environment, can predict whether the environment contains a zinc site and outputs a probability density over the input space which can be used to maximize the probability of metal binding by varying the environment through rotamer optimization. Several applications for the model related to protein structure prediction, drug discovery and protein design will be presented.

(1) Andreini, C.; Bertini, I.; Cavallaro, G.; Holliday, G. L.; Thornton, J. M. Metal Ions in Biological Catalysis: From Enzyme Databases to General Principles. J. Biol. Inorg. Chem. JBIC Publ. Soc. Biol. Inorg. Chem. 2008, 13 (8), 1205–1218. https://doi.org/10.1007/s00775-008-0404-5.
(2) Der, B. S.; Edwards, D. R.; Kuhlman, B. Catalysis by a De Novo Zinc-Mediated Protein Interface: Implications for Natural Enzyme Evolution and Rational Enzyme Engineering. Biochemistry 2012, 51 (18), 3933–3940. https://doi.org/10.1021/bi201881p.
(3) Bozkurt, E.; Perez, M. A. S.; Hovius, R.; Browning, N. J.; Rothlisberger, U. Genetic Algorithm Based Design and Experimental Characterization of a Highly Thermostable Metalloprotein. J Am Chem Soc 2018, 140 (13), 4517–4521. https://doi.org/10.1021/jacs.7b10660.
(4) Guffy, S. L.; Der, B. S.; Kuhlman, B. Probing the Minimal Determinants of Zinc

Binding with Computational Protein Design. Protein Eng. Des. Sel. 2016, 29 (8), 327–338. https://doi.org/10.1093/protein/gzw026.
(5) Skalic, M.; Varela-Rial, A.; Jiménez, J.; Martínez-Rosell, G.; De Fabritiis, G. LigVoxel: Inpainting Binding Pockets Using 3D-Convolutional Neural Networks. Bioinformatics 2019, 35 (2), 243–250. https://doi.org/10.1093/bioinformatics/bty583.
(6) Shroff, R.; Cole, A. W.; Diaz, D. J.; Morrow, B. R.; Donnell, I.; Annapareddy, A.; Gollihar, J.; Ellington, A. D.; Thyer, R. Discovery of Novel Gain-of-Function Mutations Guided by Structure-Based Deep Learning. ACS Synth. Biol. 2020. https://doi.org/10.1021/acssynbio.0c00345.

A high throughput assay to measure stability of proteins in vivo

Ingemar André

Biochemistry and Structural Biology, Lund University

TBA

Poster Abstracts

SeqPredNN: a neural network that predicts protein sequences that fold into specified tertiary structures

F. Adriaan Lategan, Hugh -G. Patterton
Center for Bioinformatics and Computational Biology

Finding amino acid sequences that assume a target structural conformation is an important step in the de novo protein design process. Recent successes in protein structure prediction suggest that machine learning may be a valuable tool to address this inverse folding problem. We demonstrate that even the simplest neural network architectures can accurately assign sequences to protein structures. We present a feed-forward neural network model (SeqPredNN) that predicts the identity of amino acids in a protein structure using only the relative positions, orientations, and backbone torsional angles of nearby residues in the PDB structure. This method accurately predicts 29.4% of the amino acids in the test set. More importantly, structures predicted by both AlphaFold and RoseTTAFold suggest that the sequences generated by SeqPredNN have folded structures that are very close to the crystal structure of the native sequences. We compare the predicted structures for sequences generated by the graph neural network ProteinSolver by Strokach et al. and the convolutional neural network ProDCoNN by Zhang et al. We show that SeqPredNN produces similar results to these methods, despite the comparative simplicity of its neural network architecture. SeqPredNN is available at https://github.com/falategan/SeqPredNN

How many knotted proteins are within the human proteome?

Agata P. Perlinska, Wanda Niemyska, Bartosz A. Gren, Marek Bukowicki, Szymon Nowakowski, Pawel Rubach, Joanna I. Sulkowska

Interdisciplinary Laboratory of Biological Systems Modelling Centre of New Technology, Warsaw, Poland

The fact that proteins can have their chain formed in a knot is known for almost 30 years. However, as they are not common, only a fraction of such proteins is available in the PDB. It was not possible to assess their importance and versatility up until now because we did not have access to the whole proteome of an organism, let alone a human one. The arrival of efficient machine learning methods for protein structure prediction, such as AlphaFold and RoseTTaFold, changed that. We analyzed all proteins from human proteome (over 20 000) in search for knots and found them in less than 2% of the structures. Using a variety of methods, including homolog search, clustering, quality assessment and visual inspection, we determine the nature of each of the knotted structures and classify it as either knotted, potentially knotted or an artifact (see our database available at https://knotprot.cent.uw.edu.pl/alphafold).

Predicting changes in binding energy upon mutations using Flex ddG with talaris2014 and beta_nov16

Aleksandra Panfilova, Johanna K. S. Tiemann, and Amelie Stein
Department of Biology, University of Copenhagen

Single point missense mutation can be sufficient for the loss of function. It can be caused by destabilization of the protein and its degradation or by loss of activity or interactions. From the clinical perspective, it would be beneficial both to know nature of disorder and to predict the effect of a specific variant. To separate loss of function caused by destabilization from the one caused by loss of interaction, tools for predicting binding ddG such as Flex ddG are required along with monomeric ddG predictors. The Rosetta Flex ddG protocol by Barlow et. al. predicts a change in binding energy upon sequence mutation on a protein structure in the interface of a protein complex. To assess how well it works in our hands, we repeated the original benchmark and extended it by comparing it to more recent score function beta_nov16.

Towards generalizable prediction of thermo-stability for scFvs using machine learning on sequence and structure features

Ameya Harmalkar, Kathy Y. Wei
Institute of Biochemistry, Technical University of Graz, Austria, Technical University of Graz, Austria

In the last three decades, the appeal for monoclonal antibodies (mAbs) as therapeutics has been steadily increasing as evident with FDA's recent landmark approval of the 100th mAb. However, unlike mAbs that bind to single targets, multispecific biologics with their single-chain variable fragment (scFv) modules have the advantage of engaging distinct targets and have garnered particular interest. Despite their exquisite specificity and affinity, the relatively poor thermostability of these scFv modules often hampers their development as a potential therapeutic drug. In recent years, engineering antibody and multispecific biologics sequences to enhance their stability by mutations has gained considerable momentum. As experimental methods for antibody engineering are time-intensive, laborious, and expensive, computational methods serve as a fast and inexpensive alternative to conventional routes. Here, we show two machine learning methods to classify thermostable scFv variants from sequence. Both these models are trained over temperature-specific data (TS50 measurements) derived from multiple libraries of scFv sequences and measured by phage-display. In this work, we show that one model generalizes better than the other. Further, we demonstrate that these models trained on TS50 data could identify 18 residue positions and 5 identical amino-acid mutations for an independent published mAb dataset. Further, transferring such models for alternative physicochemical properties of antibodies can have potential applications in optimizing large-scale production and delivery of multispecific biologics.
Keywords: multispecific biologics design, thermostability prediction, machine learning

---

Proteome-Based Immunoinformatic Design of Multi-Epitope Vaccine Candidate Against Trypanosoma brucei gambiense

Ammar Usman Danazumi*, Salahuddin Iliyasu Gital, Salisu Idris, Lamin Bs Dibba, Emmanuel Oluwadare Balogun and Maria Wiktoria Górna*

Warsaw University of Technology, University of Warsaw

Human African trypanosomiasis (HAT) is a neglected tropical disease that is caused by flagellated parasites of the genus Trypanosoma. HAT imposes a significant socio-economic burden on many countries in sub-Saharan Africa and its control is hampered by several drawbacks ranging from the ineffectiveness of drugs, complex dosing regimens, drug resistance, and lack of a vaccine. Despite more than a century of research and investigations, the development of a vaccine to tackle HAT is still challenging due to the complex biology of the pathogens. Advancements in computational modeling coupled with the availability of an unprecedented amount of omics data from different organisms have allowed the design of new generation vaccines that offer better antigenicity and safety profile. One of such new generation approaches is a multi-epitope vaccine (MEV) designed from a collection of antigenic peptides. A MEV can stimulate both cellular and humoral immune responses as well as avoiding possible allergenic reactions. Herein, we take advantage of this approach to design a MEV from conserved hypothetical plasma membrane proteins of T. brucei gambiense, the trypanosome subspecies that is responsible for the west and central African forms of HAT. The designed MEV is 402 amino acids long (41.5kDa). It is predicted to be antigenic, non-toxic, to assume a stable 3D conformation, and to interact with a key immune receptor. In addition, immune simulation foresaw adequate immune stimulation by the putative antigen and a lasting memory. Therefore, the designed chimeric vaccine represents a potential candidate that could be used to target HAT.

Benchmarking Rosetta's protein stability model using new high-throughput experiments

Andres Lira, Dr. Kotaro Tsuboyama, Dr. Gabriel Rocklin
Northwestern Univeristy, Rosetta Commons Post Baccalaureate

Predicting protein stability is important in protein design to produce effective therapeutics and understand diseases related to unstable proteins. Rosetta is a computational model used to predict the stability of proteins, among other applications. Previous tests of Rosetta's stability model have been limited to comparisons with 1,000 $\Delta\Delta$G measurements. To expand this dataset, we recently measured millions of $\Delta\Delta$G values using a new high-throughput proteolysis-based assay, cDNA assay. We hypothesize that using this larger dataset will more accurately quantify the accuracy of the Rosetta protein stability model. We have quantified the accuracy of Rosetta by determining Pearson's correlation between Rosetta scores and cDNA assay data. We have found the correlation of lab-designed proteins to have a median of -0.52 +/- 0.14 and natural proteins to have a median of -0.52 +/- 0.11. Our high-throughput accurate measurements will allow rapid improvement and development of future protein stability prediction models.

Identification of potential inhibitors for the antimalarial target PfPMX from African databases of Natural Products

Cheickna Cisse[1], Oudou Diabate, Mamadou Sangare, Jeffrey Shaffer, Seydou Doumbia, Mamadou Wele

African Center of Excellence in Bioinformatics at the Université of Sciences, Techniques and Techniques of Bamako, USTTB, Institute Pasteur of Tunis, Tunisia

Objective: The objective of this study was to model the structure of the multistage drug target Plasmepsin X from Plasmodium falciparum and to find out the potential inhibitors from African databases of Natural Products.
Methods: The model was built based on multiple templates approach implemented in Modeller. The crystallographic structure of the protein (Code PDB: 6ORS) and the model predicted by AlphaFold (code UniProt Q8IAS0) was used as templates. A high throughput virtual screening was carried out via Autodock vina by using the following African databases of Natural Products: NANPDB, EANPDB, AfroDb, and SANCDB.
Results: The results presented the accurate and validated model of the Plasmepsin X built by rational modeling for structure-based virtual screening. A total of 8690 compounds were screened among which 68 have shown high affinity (docking score $\leqslant$ -9 kcal/mol) with the protein. The detail of interactions showed that those compounds interact with key residues of the protein, making them promising inhibitors.
Conclusion: The findings of this study identified promising inhibitors to be added to the list of potential drugs against Plasmodium falciparum, the main agent of Malaria.

Optimizing Computational PROTAC Design for Targeted Protein Degradation

Chelsea Shu, Shourya Sonkar Roy Burman, Eric Fischer
Harvard University Graduate School of Arts and Sciences, Suite 350

New pharmacological modalities that repurpose the ubiquitin degradation system possess the ability to degrade proteins once thought impossible to target. Proteolysis-targeting chimeras (PROTACs) are leading examples. These hetero-bifunctional molecules contain two binding moieties covalently connected by a linker: an E3 binding ligand and a small molecule that specifically binds to a protein of interest (POI). When the E3 ligand binds to CRBN, a substrate receptor on the CRL4 E3 ubiquitin ligase, the PROTAC induces the formation of a ternary complex between the POI and E3 ligase2, activating the ubiquitin ligase system for degradation. Unlike traditional inhibitors, which are occupancy-dependent, degradation by PROTACs is event-driven, meaning sub-stoichiometric dosage of compound will lead to sustained loss of protein activity. PROTAC therapeutics hold immense potential for treating an expansive range of illnesses, including cancer and inflammatory illnesses. However, this potential is currently limited as only 2 PROTACs have entered clinical trials due to the time-consuming process of PROTAC design and testing. I will create an *in-silico* program that can efficiently predict the conformations of PROTAC ternary complexes. The program will be comprised of three quintessential steps that will recreate the 10 PROTAC ternary complexes cited in literature.

Using Rosetta to increase solubility and prevent aggregation of M. amurensis seed leucoagglutinin

Eva Rajh, Helena Gradišar, Ajasja Ljubetič, Neža Omersa, Roman Jerala
National Institute of Chemistry, Department of Synthetic Biology and Immunology

Solubility of a protein is an important parameter for all protein related applications and its intrinsic solubility is primarily defined by amino acids on the protein's surface. Increased negative charge of the protein's surface is strongly correlated with its solubility (1). Maackia amurensis seed leukoagglutinin (MAL) is a sialic-acid binding lectin that recognizes terminal sialic acid of alpha 2,3 syalillactose and is a useful tool for detecting sialoglyco-conjugate on the surface of human cancer cells (2). We would like to use MAL to functionalise coiled coil protein origami structures (3) .Expressing MAL in E. coli resulted in insoluble and aggregated protein. Since aggregation is another protein parameter modulated with surface charge (4), we used Rosetta netcharge energy term to design supercharged MAL proteins that would show increased solubility and decreased tendency for aggregation compared to the wild type (5). We generated candidates with a range of negative charges from -13 (close to wild type) to -29. We did not redesign the residues in and near the ligand binding site to preserve sugar binding. Redesigning the surface also removed a probable dimerization interface. Six constructs with different net negative charge were expressed in E. coli. Supercharging noticeably increased the solubility of MAL protein, however the majority of the purified protein was still aggregated. Using surface plasmon resonance we were unable to detect binding of the sugar to MAL, which we attributed to missing glycosylation of MAL (2).

1. Kramer, R. M., Shende, V. R., Motl, N., Pace, C. N. & Scholtz, J. M. Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility. Biophysical Journal 102, 1907 (2012).
2. Kim, B. S., Hwang, H. S., Park, H. & Kim, H. H. Effects of selective cleavage of high-mannose-type glycans of Maackia amurensis leukoagglutinin on sialic acid-binding activity. Biochimica et Biophysica Acta (BBA) - General Subjects 1850, 1815–1821 (2015).
4. Raghunathan, G. et al. Modulation of protein stability and aggregation properties by surface charge engineering. Molecular BioSystems 9, 2379–2389 (2013).
5. Supercharge - redesigning protein surfaces with high net charge.
https : *//www.rosettacommons.org/docs/latest/application*$_d$*ocumentation/design/supercharge.*

Predicting the molecular and functional impacts of genetic variants related to obesity by a personalized structural biology pipeline.

Felipe Engelberger, Georg Künze, Jens Meiler
Institute for Drug Discovery, Leipzig University Medical School

The genome of each human contains a unique set of genetic variants. Recently, the development of cheaper sequencing technologies has made available massive amounts of genomic data. Genetic variants are at the root of multiple diseases, one particular example is obesity where recently the clinical relevance and possible therapeutic potential of genetic variants in adipokine genes including leptin, chemerin, vaspin, and neuregulin-4 has gained attention. Loss-of-function mutations in these genes may be linked to development of obesity, adverse fat distribution, insulin resistance, and type 2 diabetes. In the present work we leverage a previously created Personalized Structural Biology (PSB) pipeline to study non-synonymous variants in the Leptin gene. The pipeline aims to streamline the assessment of variants of unknown significance (VUS) by predicting: (1) thermodynamic destabilization of the corresponding protein structure, (2) functional importance of the VUS by means of 3D spatial clustering of known disease-causing and benign variants with respect to VUS. For structural and energetic analyses, the pipeline retrieves structures from PDB, SWISS-MODEL and Alphafold2 databases and scores the likelihood of the VUS for being disease-causing using statistical predictors like PATHprox. In this study, we applied the PSB pipeline to the investigation of a particular variant in leptin, a monogenic obesity related protein, and its impact on leptin receptor interaction. We found that the variant Q84K is spatially close to other disease causing variants, leading to a high PATHprox score, which indicates that the Q84K variant is likely pathogenic. Furthermore, after visual inspection of the neighborhood of the mutated position we find that a salt bridge between the Q84 sidechain and the backbone of the neighboring L104 is likely to be disrupted in the Q84K variant. The disruption of the aforementioned salt bridge is confirmed by measuring their distance in a molecular dynamics simulation. Finally, RMSD and RMSF analysis of the variant with respect to the mutant suggest that losing the salt bridge may impact protein stability by making the loop containing the mutation more conformationally flexible/diverse.

Helix-Helix Loop Closure Quality

Florian Wieser

Institute of Biochemistry, Graz University of Technology

Helix bundles constitute a commonly observed protein fold in nature that plays vital functional roles in protein-protein interactions as well as protein small molecule and protein DNA interactions. Due to their functional flexibility, their structural stability as well as their well-defined secondary structure, they are promising candidates to be used as scaffolds for functional protein designs. Helix bundles are composed of two to several parallel or antiparallel alpha helices connected by structurally poor defined loop regions. For *de novo* protein design purposes, these loop regions can be potential sources of decreased protein stability and subsequently impaired folding capability, if not designed carefully. Judging the quality of all loops present in a set of helix bundle designs manually emerges as cumbersome, not at least because of the unstructured nature of loops. In a preliminary study, we established a protocol, implemented as a mover in RosettaScripts, which captures five quality features of loops within helix bundles. We benchmarked our quality feature-score on all helix bundles of the Protein Data Bank as well as on a set of small *de novo* designed helical structures. Currently, we are working on a follow-up study based on Machine Learning with the aim to establish a single metric to assess the quality of helix-helix loops and potentially loop connections between secondary structure elements in general.

Specific Protein-Protein Interfaces: Towards multi-component modular protein systems

Frances Gidley, Fabio Parmeggiani
University of Bristol, School of Chemistry

Protein assemblies facilitate many essential roles in biological systems, from structural support to communication and catalysis. Designed self-assembling modular protein structures with specific geometric and spatial properties can be applied to novel biomaterial needs, including in the influencing of cell behaviour through the organised display of extracellular signal molecules. Use of computational design and modelling methods to design novel protein-protein interfaces and large protein scaffold assemblies will enable production and design of specific, functionalisable, multi-component assemblies and production of tuneable nanomaterials for bio applications and beyond. A set of novel orthogonal protein dimers are described here, which show good solubility and specificity. These novel dimeric interfaces are modular building blocks, and larger designs capable of self-assembly into larger protein assemblies are currently undergoing experimental characterisation.

(1) Anand-Achim, N.; Eguchi, R. R.; Mathews, I. I.; Perez, C. P.; Derry, A.; Altman, R. B.; Huang, P.-S. Protein Sequence Design with a Learned Potential. bioRxiv 2021, 2020.01.06.895466. https://doi.org/10.1101/2020.01.06.895466

Modeling of the Wrap1 Protein from Nostoc punctiforme

Jędrzej Kubica[1,2], Dominik Gront[2], and Stanisław Dunin-Horkawicz[1,3]

[1] Laboratory of Structural Bioinformatics, Institute of Evolutionary Biology, University of Warsaw, Poland [2] Laboratory of Theory of Biopolimers, Faculty of Chemistry, University of Warsaw, Poland [3] Department of Protein Evolution, Max Planck Institute for Biology, Tübingen, Germany

Multicellularity is a process strictly coupled to the mechanisms of programmed cell death (PCD). A homology between well-studied eukaryotic PCD proteins (human Apaf-1, *Caenorhabditis elegans* Ced-4) and yet not fully characterized prokaryotic PCD-like proteins has been reported [1]. Wrap1 is a transmembrane protein, which is a component of a putative PCD apparatus in a multicellular cyanobacterium Nostoc punctiforme. It is equipped with a highly-repetitive β- propeller domain, which suggests its possible antibody-like role [1]. Upon activation, Apaf-1 and Ced-4 oligomerize into heptamers and octamers, respectively, to form apoptosomes – protein complexes triggering signaling pathways in apoptosis and innate immunity. Due to homology, Wrap1 is also expected to oligomerize in a similar fashion. This study has been focused on the NTPase domain of Wrap1 protein and the prediction of its possible oligomerization state with the aid of Rosetta and AlphaFold2 methods.

[1] Dunin-Horkawicz S, Kopec KO, Lupas AN. Prokaryotic ancestry of eukaryotic protein networks mediating innate immunity and apoptosis. J Mol Biol. 2014;426(7): 1568-1582.

Analysis of common buffer molecules' parameters in ligand-protein complexes

Joanna M. Macnar, Dariusz Brzeziński, Dominik Gront

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

Structural information about ligand-macromolecule complexes is key for biomedical sciences, especially for structure-based drug design and structural bioinformatics. Most of the experimental information about macromolecules complexed with ligands is coming from X-ray crystallography. The accessibility of synchrotron sources with modern appliances and software has resulted in the availability of approximately 190,000 structures deposited in the Protein Data Bank (PDB). Alas, a small number of the crystal structures available in the PDB are of suboptimal quality, including some with poorly identified and modeled ligands in protein-ligand complexes. BioShell 3.0 is an advanced structural bioinformatics toolkit that includes the analysis of ligand-protein complexes. By combining a graph-theoretical approach, kernel density estimators, bioinformatics methods, and chemical knowledge, we present an analysis verifying the accuracy of HEPES and MES molecules, frequent components of crystallization buffers, selected from PDB deposits. This analysis will lead to a better refinement of ligand-protein complexes.

Structure-based prediction of HDAC6 substrates reveals determinants of promiscuity and detects new potential substrates

Julia K. Varga [1*], Kelsey Diffley [2*], Katherine R. Welker Leng [2], Carol A. Fierkeb [3], Ora Schueler-Furman [1]

[1] Department of Microbiology and Molecular Genetics, Faculty of Medicine, Hebrew University, Israel

[2] Department of Chemistry, University of Michigan, United States

[3]Department of Biochemistry, Brandeis University, United States

[*] Equal contribution

Histone deacetylases play important biological roles well beyond the deacetylation of histone tails. In particular, HDAC6 is involved in multiple cellular processes such as apoptosis, cytoskeleton reorganization, and protein folding, affecting substrates such as α-tubulin, Hsp90 and cortactin proteins. We have applied a biochemical enzymatic assay to measure the activity of HDAC6 on a set of candidate unlabeled peptides. These served for the calibration of a structure-based substrate prediction protocol, Rosetta FlexPepBind, previously used for the successful substrate prediction of HDAC8 and other enzymes. A proteome-wide screen of reported acetylation sites using our calibrated protocol together with the enzymatic assay provide new peptide substrates and avenues to novel potential functional regulatory roles of this promiscuous, multi-faceted enzyme. In particular, we propose novel regulatory roles of HDAC6 in tumorigenesis and cancer cell survival via the regulation of EGFR/Akt pathway activation. The calibration process and comparison of the results between HDAC6 and HDAC8 highlight structural differences that explain the established promiscuity of HDAC6.

CABS-flex as a de novo structure prediction tool for cyclic peptides

Karol Wróblewski, Aleksandra Badaczewska-Dawid, Mateusz Kurciński, Sebastian Kmiecik

Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw

The structural modeling of peptides is an important problem for the discovery of new drugs and deeper understanding of the molecular mechanisms of life. Here we present a novel multiscale pipeline for the molecular structure prediction of cyclic peptides. The protocol consists of two main stages: coarse-grained simulations using CABS-flex standalone package and Modeller-based protocol for all-atom reconstruction of cyclic peptides. We evaluated the proposed protocol on a benchmark consisting of peptides with cyclization through backbone and disulphide bonds, mostly from the Cyclotides family. We also demonstrated a comparison between a novel CABS-flex protocol and PEPFOLD 2.0 method on a set of peptides with disulfide cyclization. Initial results indicate that further improvements in all-atom reconstruction is needed. Our current efforts are focused on integrating CABS-flex modeling with all-atom reconstruction and refinement done by ROSETTA. We plan to release this novel protocol as a part of CABS-flex 2.0 server service and CABS-flex standalone package in order to assist with in silico design and study of cyclic peptides drugs.

Bayesian Modeling and Analysis of Dose-Response Relationships

Madeline J. Martin, Matthew J. O'Meara Ph.D.
University of Michigan, Ann Arbor Campus

A central strategy in protein structure vs. function studies is to model how molecular systems respond to changing conditions such as temperature or the presence of a binding partner and compare against experimental data. To draw credible inferences, it is important to take into account uncertainty in both the model predictions and experimental data, however researchers typically only fit maximum likelihood models that only give point estimates, e.g. by fitting sigmoidal dose-response models through Prism or the drc R package. Alternatively, Bayesian modeling provides a sound statistical framework to quantify uncertainty by producing posterior probability distributions over the model parameters. To facilitate the broader adoption of Bayesian modeling for protein structural vs. function and dose-response studies, I will present a new BayesPharma R package that builds on recent computational and methodological advances in the Bayesian modeling field. To demonstrate, I will show a re-analysis of a Kappa Opioid Receptor (KOR) electrophysiology assay and show how Bayesian models are better able to discriminate significant and non-significant differences between drugs compared to standard maximum likelihood fits. Then I will describe a plan to facilitate the broader practical adoption of rigorous Bayesian modeling in the Rosetta community. This involves 1) enhancing the Rosetta scientific benchmarks to better account for modeling and experimental uncertainty and 2) developing case studies, teaching material, and workshops to train Rosetta users and researchers in the broader computational biophysics field.

Shape based protein design

Mads Jeppesen, Ingemar Andre
Lund University, Centre for Molecular Protein Science

Protein assemblies are abundant in nature and carry out much of the functional complexity of nature. They are useful architectures for protein design as they open the door for a plethora of new functionalities and technologies that could benefit humanity. To date, much success has been achieved in designing assemblies with different architectures. However, designing more complex structures with multiple interfaces is still an outstanding challenge. One major obstacle in designing multiple interfaces is the simultaneous preservation of the shape complementarity across all interfaces. To meet this challenge, I present an in-progress method for designing protein assemblies from a shape-based paradigm. The approach has 3 steps: 1. Defining excellent shapes that would fit into a target assembly. 2. Extracting proteins from a database with similar shapes and aligning them into the assembly. 3. Redesign of the extracted protein interfaces. We are currently testing our methodology on experimental designs of native like protein capsid where each subunit has on the order of 3-4 unique interfaces.

CNN method for protein backbone reconstruction

Maksymilian Głowacki

Inter-faculty Individual Studies in Mathematics and Natural Sciences College, University of Warsaw

Protein backbone atoms provide an invaluable information such as hydrogen bonding network and secondary structure. Reconstruction of protein main chain atomic positions is the first step in reconstructing full atom conformation from a coarse grained model, which often provides only C-alfa coordinates.
We propose a novel method based on prediction of lambda dihedral angles with convolutional neural networks (CNNs). Our CNN model makes use of precalculated features for every residue, including distances between C-alfa atoms in a local neighbourhood, 3-state secondary structure prediction and number and energy of hydrogen bonds formed. Our study provides an alternative and accurate method to predict lambda dihedral angles, which may promote protein structure prediction and further development of multiscale modelling approaches.

Crystallizing Novel Proteins for Blind Testing of Improved Structure Refinement Tools in Foldit

Mark Bray[1] , Andreas Petrides [2] , Firas Khatib [2] , Scott Horowitz [1]

[1]Knoebel Institute for Healthy Aging, University of Denver, Denver, CO
[2]Department of Computer & Information Science, University of Massachusetts, Dartmouth, MA

The lack of knowledge of many proteins' structures currently limits biological research. There exists a need for novel approaches to efficiently and accurately solve the structures of new proteins. Crowdsourcing via the protein-folding game Foldit is a promising method for citizen- scientists to contribute to protein structure determination, as previous research has demonstrated the ability for citizen scientists to improve model building accuracy. A major stumbling-block in Foldit development is the ability to perform blind test puzzles. To facilitate blind tests of new structure-solving Foldit features, we will crystallize proteins with no presently known sequence homology, as verified by PDB, HHsearch, PSI-BLAST, and AlphaFold2 searches. We will express and purify proteins of interest in E. coli, then screen for crystal hits or refine preliminary crystals by vapor diffusion using both sitting and hanging drop methods. Data will be collected at synchotron data sources. Because our proteins have no sequence homologs in the PDB, we anticipate the need to perform experimental phasing using selenomethionine incorporation. At present, we have successfully expressed and purified one protein, and are starting crystallization trials. The development and testing of improved structure refinement tools in Foldit will allow for faster, more efficient, and collaborative protein structure determination for the benefit of many biological scientific disciplines.

Multiple alignment method for identification of catalytic microenvironments in enzymes

Marko Babić, Daniela Kalafatović
University of Rijeka, Department of Biotechnology and Drug Development

By analyzing the primary sequence composition of the EC 3.1 class of hydrolase enzymes, we aim to accelerate the discovery of short catalytic peptides. The analysis showed highly conserved catalytic sites with distinct positional patterns and highlighted three features critical for enzyme efficiency, being (i) the catalytic residues determined by mechanism studies, (ii) the spatial geometry of mentioned residues and (iii) the conserved chemical properties of catalytic microenvironments.While classical multiple sequence alignment (MSA) is excellent for analyzing evolutionary traits, the comparison of significantly dissimilar enzyme domains might pose problems, especially when aligning small functional segments shared by these domains. For example, patterns in neighboring sequences near catalytic residues, called catalytic microenvironments, might be missed across superfamily analyses due to proteins being generally too dissimilar. Another downside is that alignments focus on side chain identities and may miss atom positioning or even a general chemical property conservation. For example, the main chain oxygen of tryptophan might have catalytic function equal to a side chain oxygen of aspartate and the alignment will interpret the position as unpreserved. Current structural alignment and structure-sequence alignment combinations focus mainly on α-carbon positions.To circumvent these issues and make enzyme function analysis more informative we analyzed catalytic microenvironments of ester hydrolases with a catalytic triad mechanism. To create these microenvironments having 32 amino acid (AA) lengths, 16 amino acids were extrapolated from each side of the catalytic residues responsible for catalysis. Next, we performed structural alignment using Pymol centered on the positions of the five catalytic residues: catalytic nucleophile, base, acid, and two oxyanion hole members. The five 32-length AA sequences per protein were taken from FASTA files from RCSB PDB database using filters for EC and CATH. Redundant sequences were removed and the structures were stripped to include only chain A of the protein to remove bias. Similarly, PDB files were extracted for structural alignment, where the additional removal of heteroatoms was performed to prevent interference with the alignments. In addition, classic ClustalW and VMDs structural alignments were performed.Preliminary results show clear distinctions in superfamily-specific conservation and residues of direct catalytic importance. The data additionally showed reactive atom positioning and chemical property conservation missed by classical methods, elucidating the potential for a different approach for identifying important aspects of enzymatic function. Peptide sequences built upon these studies might be evaluated for potential catalytic activities alone or inserted in scaffolds.

Influence of two novel mutations in MLASA disease on YARS2 - tRNA complex

Mateusz Fortunka, Agata Perlińska, Adam Stasiulewicz, Ya-Ming Hou, Joanna Sułkowska

Interdisciplinary Laboratory of Biological Systems Modelling Centre of New Technology, Warsaw, Poland

Recently found two novel mutations of YARS2 protein cause severe symptoms of rare genetic disease MLASA. The idea of the research was to explain the connection between two indirect mutations and the enzymatic reaction stop. YARS2 tRNA binding sites lacking in crystal structure were modelled using homologic proteins in complex with tRNA. Then, the homologic tRNA from T. thermophilus was humanized and docked to YARS2. Numerous models obtained in rigid docking were numerically analysed. The best of them were simulated using molecular dynamics, which led to choose the best structure. It was used to perform longer simulations of mutated and non-mutated versions of the complex. The last step was analysing the data by the dynamical network analysis, which led me to propose mechanism of mutation influence on the catalysis. It turns out that breaking the symmetry in two protein chains by mutations leads to signal dispersion after tRNA binding, so that its 3' end cannot enter the active site. During the project, there were created tools to analyse big sets of data coming from docking and dynamical network analysis.

Non-redundant subsets of protein structures from a sequence-based hierarchical clustering method

Mateusz Skłodowski[1], Joanna M. Macnar[2,3], Maksymilian Glowacki [2], Dominik Gront[2]

[1] Institute for Drug Discovery, Leipzig University Medical School, Liebigstraße 27, 04103 Leipzig,Germany

[2] Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-089 Warsaw, Poland

[3],College of Inter-faculty Individual Studies in Mathematics and Natural Sciences, University of Warsaw, Stefana Banacha 2C, 02-089 Warsaw, Poland

Proteins have been the subject of increased scientific research for many years. Especially recently, the number of published structures has been increasing significantly [1]. However, this increment is not regular in terms of sequence diversity, so that some protein families are much better understood than others. A consequence of this phenomenon is the overrepresentation of certain motifs in the structural statistics of databases. One possible solution to this problem is to include only parts of protein groups in the statistics. For this purpose, representatives of whole groups are selected from a family of proteins to form a new database of structures called "nonredundant" [2]. The second possible approach is to introduce weights to take into account the significance of the contribution of a given structure to the statistics [3]. This allows the removal of the aforementioned biases while preserving unique sequence motifs. This approach was adopted in our work. To obtain the weights, the identity of the structures was checked using sequence alignment. Two methods were used for this purpose: BLAST [5] based on heuristic local sequence matching and HMMER [6] using hidden Markov models (HMM). Then, using cluster analysis methods [6], groups of proteins showing sequence similarity were created. A representative is selected from each grouping and its comparison with each protein in that set is stored using weights. The results obtained were divided into three subgroups depending on the sequence similarity cut-off value. Analysis of the data shows that the HMM path gives on average 7% more clusters than PISCES database [7] and 22.5% more than BLAST analysis path. Based on our results, we can conclude that the HMM pathway is the most accurate method for creating non-redundant subsets of proteins among the methods studied. Further development on the HMM pathway will take place in the Libra database created for this purpose.

[1] H. M. Berman et al., "The Protein Data Bank," Nucleic Acids Research. 2000, doi: 10.1093/nar/28.1.235.

[2] W. C. Barker et al., "The PIR-International Protein Sequence Database," Nucleic Acids Research. 1999, doi: 10.1093/nar/27.1.39.

[3] C. Yanover et al., "Redundancy-weighting for better inference of protein structural features," Bioinformatics, 2014, doi: 10.1093/bioinformatics/btu242.

[4] S. Altschul, Stephen F., et al. "Basic local alignment search tool." Journal of molecular biology 215.3 (1990): 403-410.

[5] Durbin, Richard, et al. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press. 1998, pp. 101-134, doi:10.1017/CBO9780511790492.006

Highly flexible docking of cyclic peptides to proteins using CABS-dock and Rosetta refinement

Mateusz Zalewski, Aleksandra E. Badaczewska-Dawid, Mateusz Kurciński, Sebastian Kmiecik

Biological and Chemical Research Centre, Faculty of Chemistry, University of Warsaw

The number of peptide-based therapeutics is growing continuously and it is expected to increase even more in coming years. Recently a wide attention has been drawn to cyclic peptides as potential modulators of biomolecular interactions with improved biological activity. Only a few methods exist that enable molecular docking of cyclic peptides, however with some limitations. Here, we present a protocol for a flexible docking of cyclic peptides. Method is based on the combination of well-established tool for protein-peptide docking – the CABS-dock, and Rosetta refinement. The proposed protocol was evaluated on a set of 38 cyclic peptide complexes. Provided results show that the combination of CABS-dock with Rosetta refinement may be an effective way for docking not only linear, but also cyclic peptides.

The Repeating, Modular Architecture of the HtrA Proteases

Matthew Merski

University of Warsaw, Department of Chemistry

A conserved, 26 residue sequence
[AA(X2)[A/G][G/L](X2)GDV[I/L](X2)[V/L]NGE(X1)V(X6)]
and corresponding structure repeating module was identified within the HtrA protease family using a non-redundant set (N=20) of publicly available structures. While the repeats themselves were far from sequence perfect they had notable conservation to a statistically significant level with three or more repetitions identified within one protein at a level that would be expected to randomly occur only once per 1031 residues. This sequence repeat was associated with a six stranded antiparallel *b*-barrel module, two of which are present in the core of the structures of the PA clan of serine proteases, while a modified version of this module could be identified in the PDZ-like domains. Automated structural alignment methods had difficulties in superimposing these *b*-barrels but use of a target human HtrA2 structure showed that these modules had an average RMSD across the set of structures of less than 2 $\mathring{A}$ (mean and median). Our findings support Dayhoff's hypothesis that complex proteins arose through duplication of simpler peptide motifs and domains.

Structure prediction and modeling of the *lvaD* gene expression product contained in *lva* operon of *Pseudomonas putida*

Mikołaj Iwan[1], Jędrzej Kubica[2,3]

[1] Department of Environmental Microbiology and Biotechnology, Institute of Microbiology, Faculty of Biology, University of Warsaw [2] Laboratory of Structural Bioinformatics, Institute of Evolutionary Biology, University of Warsaw, Poland [3] Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Poland

Levulinic acid (LA) is a compound of growing importance since it was designated by the US Deparment of Energy as one of 12 key sugar-derived platform chemicals that can be produced from biomass. LA can be further transformed to 4-hydroxyvalerate (4-HV) via microbial metabolism. 4-HV is similarly important due to being a monomer for polyhydroxyalkanoates (PHA) production which are emerging as environmentally friendly alternatives to synthetic polymers. Rand et al. (2017)[1] presented an exhaustive analysis of *lva* operon present in *Pseudomonas putida* KT2240 which is responsible for LA catalysis. The subject of this work is predicting and modelling the structure of parts of this metabolic pathway with hopes of better understanding the intricacies of LA catalysis. Expanding the knowledge about this process opens up opportunities for modifying the cell infrastructure involved in this process to improve yield and lower energy demand for producing the very valuable chemical which is 4-hydroxyvalerate.

1.Rand, J. M. et al. A metabolic pathway for catabolizing levulinic acid in bacteria. Nat Microbiol 2, 1624–1634 (2017).

Machine Learning Tool for Structural Bioinformatics

Mohammad N. Saqib.
University of Warsaw, Faculty of Chemistry

To comprehend a protein model generated using computational methods, secondary structure elements must be assigned to protein conformations. The procedure entails marking amino acid residues with the letters H (helix), E (strand), or C (coil/loop). The method becomes hard when key atoms are absent from an input protein structure, especially when only the positions of the alpha carbons are known. Several techniques have been tested and applied to this problem over the last forty years. The most recent development is the use of machine learning techniques. A novel classifier is presented in this project. This classifier assigns protein secondary structure categories using neural networks. Only C-alpha coordinates are used in this procedure. The Keras (TensorFlow) package was used to create and train the neural network. From the raw coordinates, the BioShell toolbox was used to compute the neural network's input features. When just the C-alpha trace is provided, the study's findings suggest that neural network-based methods can be utilized to successfully solve structural mapping problems. Our approach's accuracy (above 97 percent) outperformed prior approaches.

Ferritin-scaffolded nanoparticles as influenza vaccines

Niels E. J. Meijer

Lund University, Centre for Molecular Protein Science

This project comprises the design of a single-protein influenza vaccine candidate that avoids the long production times of traditional vaccines. The simple biochemistry also facilitates the inclusion of multiple viral strains for the sake of broad-spectrum immunity. The chosen approach involves linking the rotationally symmetic viral antigens haemagglutinin and neuraminidase to a similarly symmetric ferritin scaffold by means of computational protein design. The proper assembly and folding of these constructs are characterised experimentally afterwards. Other means of linking these antigens are also explored, such as incorporating a hyperstable coiled-coil region in the linker and the use of a helix-loop-helix clip-on system. Display on a self-assembling two-dimensional de novo protein lattice is also investigated.

Computer-aided drug discovery on cyclic nucleotide-gated ion channels

Palina Pliushcheuskaya, Georg Kuenze, Vasilica Nache, Sandeep Kesh, Frank Schwede

Institute for Drug Discovery, Faculty of Medicine, Leipzig University

Cyclic nucleotide-gated (CNG) ion channels are involved in signal transduction in retinal and olfactory systems. Their activation proceeds upon binding of endogenous cyclic nucleotides (cGMP, cAMP) and leads to conversion of biochemical signals into electrical stimulation. In case of retinal degeneration diseases, elevated cGMP concentrations and associated overactivation of CNG channels result in increased ion flux mostly into the rod photoreceptors, which could cause vision loss or other eye pathological conditions. The aim of this project is to design inhibitors selective for rod over cone CNG channels. For this purpose, we performed docking of cGMP-like ligands into rod and cone channel structures with the Rosetta software suite to investigate their modes of interaction. Interactions between CNG channels and their ligands were very similar for the rod and the cone structure because of their high sequence similarity. We discovered some minor amino acid differences close to the cyclic nucleotide-binding domain, which we will exploit to obtain selectivity. In addition, we are examining possibilities to design allosteric modulators by targeting different domains in the CNG-channel structures. As next step we will use structure-based drug design methods to suggest chemical modifications to CNG ligands with the aim to increase ligand interactions with the rod GNG channel structure and weaken binding to the cone CNG structure.

Prediction of double knotted protein structure (3 1 #3 1) based on AI method

Smita P Pilla, Agata P Perlinska, Mai Lan Nguyen, Szymon Niewieczerzal, Fernando Bruno da Silva, Iwona Lewandowska and Joanna I Sulkowska

Interdisciplinary Laboratory of Biological Systems Modelling Centre of New Technology, Warsaw, Poland

The high-resolution three-dimensional structure of proteins is crucial for the characterization of various functions and understanding the mechanisms of life. Double knotted proteins are the proteins that contain two knots and are the most complex knotted structures. Recently, the presence of two trefoil knots (31#31) in proteins has been discovered. Here, we focus on the double knot in the SpouT superfamily, which is formed by the methyltransferase: TrmD (established biological activity)2, and Tm1570 (unknown function). Based on a comprehensive bioinformatics approach, we aim to prove the existence of such a structure and understand the influence of knots on the biological activity of the protein. We used DeepMind's AlphaFold3 machine learning method to predict the structure of all homologous proteins from their sequence. We find they all form double knotted structures. Additionally, we used AlphaFold-Multimer to predict the multimeric protein. We found that there are at least two forms of dimer for double knotted proteins, both are different from the known TrmD and Tm1570. Furthermore, we used molecular dynamics simulations to determine the stability of the predicted dimer. This is the first occurrence of a knot type that is not a twist knot (i.e. is not the result of twisting of a loop followed by single threading). To form a 31#31 knot, the protein chain must cross the energy barrier at least twice during folding as it is pulled through the twisted loops.

References:

1. Niemyska, Wanda, et al. "AlphaKnot: Server to analyze entanglement in structures predicted by AlphaFold methods" Nucleic Acids Research [doi:10.1093/nar/gkn000].

2. Hou, Ya-Ming, et al. "TrmD: a methyltransferase for tRNA methylation with m1G37." The Enzymes 41 (2017): 89-115.

3. Evans, Richard, et al. "Protein complex prediction with AlphaFold-Multimer." BioRxiv (2021).

Gradient boosting classifier in GPCR drug discovery

Szymon Wiśniewski, Paulina Dragan, Mikołaj Mizera, Dorota Latek
Faculty of Chemistry, University of Warsaw, Poland

The number of ligand scaffolds constituting a chemical space of GPCR actives is a major determinant of the accuracy of ligand-based drug design. Currently available datasets in freely accessible databases may include thousands of unique ligand scaffolds for many GPCR receptors, e.g., chemokine receptors, but significantly limited datasets for others, e.g., glucagon receptors. Yet even the smallest curated dataset used for training offers the predictive relevance of Q 2 above 0.6 if the efficient supervised learning algorithm is used [1, 2]. Such a level of accuracy enables to distinguish most of the active ligands from non-actives in the applicability domain for the certain receptor dataset. Moreover, the receptor subtype selectivity can be distinguished by gradient boosting based on the most populated ligand datasets, e.g., for cannabinoid receptors. Here, we discussed the impact of the properties and scaffolds distribution on the accuracy of the ligand-based drug design [2] and compare it with the performance of the structure-based drug design [3, 4]. This provided the basis for the overcoming of typical problems with non-curated GPCR datasets (rough and fuzzy sets), and maladjustment of LBDD and SBDD algorithms while combining them [2].

We acknowledge National Science Centre in Poland (2020/39/B/NZ2/00584).

[1] Mizera, M.; Latek, D.; Cielecka-Piontek, J. Virtual Screening of C. Sativa Constituents for the Identification of Selective Ligands for Cannabinoid Receptor 2. Int. J. Mol. Sci. 2020, 21, 5308.
[2] Mizera, M.; Latek, D. Ligand-Receptor Interactions and Machine Learning in GCGR and GLP-1R Drug Discovery. Int. J. Mol. Sci. 2021, 22, 4060.
[3] Latek D, Rutkowska E, Niewieczerzal S, Cielecka-Piontek J. Drug-induced diabetes type 2: In silico study involving class B GPCRs. PLoS One. 2019;14(1):e0208892.
[4] Langer I, Latek D. Drug Repositioning For Allosteric Modulation of VIP and PACAP Receptors. Front Endocrinol (Lausanne). 2021;12:711906.

A generalisable pipeline for the design of multi-haem proteins for effecient electron transfer.

T. Neary, R. Anderson, F. Parmeggiani
University of Bristol, Bristol

Haem proteins represent a broadly specific class of proteins capable of diverse catalysis, including electron transfer reactions. However, to date, no designed proteins have been able to demonstrate efficient unidirectional transfer of electrons across multiple haem cofactors. Designing a protein capable of such requires precise spatial organisation to ensure minimal cofactor edge-edge distance and optimal local environmental conditions to facilitate precise electrical mid point potentials. Here we describe a method for designing repeat proteins capable of binding multiple haem groups in precise oreintations. We demonstrate that this method can be applied more generally to automate the approach to optimising and streamlining the design of arbitrary protein targets.

Modelling and Analysis of 3D structure of full length of wild and mutant KIT-kinase

Udit Gupta[1,2], Siddhi Bhutada[1,2], Harshul Diwanji[1,2], 1Debjani Dasgupta, Pramodkumar P Gupta[1*]

[1]School of Biotechnology and Bioinformatics, D Y Patil Deemed to be University, Plot 50, Sector 15, CBD Belapur, Navi Mumbai 400614, Maharashtra, India.

[2]Thadomal Shahani Engineering College, W, P. G. Kher Marg, (32nd Road, Marg, Off Linking Rd, TPS III, Bandra West, Mumbai 400050, Maharashtra, India.

[*]Corresponding author: Pramodkumar P Gupta

c-Kit, a receptor tyrosine kinase, is involved in intracellular signaling, and its mutation plays a crucial role in the occurrence and progression to numerous types of cancers. As per the data present on Uniprot database, KIT kinase does not have a complete and verified structure available, thus leading us to take the 976 amino acid sequence from the database to provide one. Gastro-Intestinal Stromal Tumor (GIST), Acute Myeloid Leukemia (AML), prostate cancer etc. are cancers being affected by the mutations in c-Kit. GIST was selected as the cancer to focus on due to the availability of robust and well-corroborated data regarding it and its linked KIT kinase mutations. Using these sequences and mutations, we began in silico structure generation. We used homology modelling (SWISS-MODEL server), fold recognition (Phyre2), and ab initio modelling (Robetta). Using GROMACS MD simulation package, all the modelled 3D structure were solvated under the SPCE water model using OPLS method. The solvated protein system was neutralized by adding desired number of Na+ Cl- ions. Later the energy minimization was performed using Steepest Descent method to attain the lowest or global minimum energy form for the stability of the protein. All the modelled 3D structures pre and post energy minimization were used for Ramachandran Plot analysis using SAVES server and ProSA based analysis was done.

From SWISS-MODEL server we have received incomplete structure due to the partial homology of the template so used. Phyre-2 and Robetta server provided us with complete 3D structure of entire length of 976 amino acid residues. From pre and post energy minimization the protein structure reported from 95% to 99% residues in the allowed regions of the Ramchandran plot, ProSA reported most of the structure in the accepted region. The energy minimization results highlighted the accepted minimum linear energy form for Potential energy, Bond, Angle and Proper-Dihedral. The stability of all the four factors reported a stable a conformation of the modelled 3D structures. Since we have used multiple approaches to predict the complete length of wild KIT kinase, we can compare the efficiency of these various prediction methods. We aim to use these structures further for MD simulation and molecular interaction study in understanding drug binding effects.

Key words: c-Kit, KIT kinase, Model, 3D structure.

Computational design of dimeric de novo heme-binding helical bundle proteins

Veronica Delsoglio

Institute of Biochemistry, Technical University of Graz, Austria,Technical University of Graz, Austria

Proteins mediate the fundamental processes of life and have been the focus of much biomedical research for the last 50 years; indeed, protein-based materials can solve a vast array of technical challenges. The naturally occurring proteins mediate essential functions like the use of solar energy to manufacture complex molecules, the ultra-sensitive detection of small molecules and of light, the conversion of pH gradients into chemical bonds and the transformation of chemical energy into work. All these functions are encoded in sequences of amino acids with extreme economy, and such sequences specify the three-dimensional structure of the proteins. If the fundamentals of protein folding and protein biochemistry and biophysics can be understood, it should become possible to design customised proteins which could address many of the important challenges that society faces. Natural evolution has produced a vast array of proteins that perform the physical and chemical functions required for life; although proteins can be reengineered to provide altered or novel functions, the utility of this approach is limited by the difficulty of identifying protein sequences that display the desired properties. Recently, advances in the field of computational protein design have shown that molecular simulation can help to predict sequences with new and improved functions (de novo protein design). Moreover the computational methodology has advanced to the point that a wide range of structures can be designed from scratch with atomic-level accuracy. The problem of the de novo protein design is that both the sequence and the exact structure of the backbone are unknown. The design calculations generally begin with large set of alternative conformations (more that 10,000) and these initial backbones can be made either by assembling short peptide fragments or by using algebraic equations to specify the geometry parametrically. De novo designs are usually experimentally characterised only if structure-prediction calculations that start from the designed sequence strongly converge on the designed structure. In order to perform the de novo protein design, as well as other computational works, Rosetta software suite is used, which includes algorithms for computational modeling and analysis of protein structures. Rosetta development began in the laboratory of Dr. David Baker at the University of Washington as a structure prediction tool, and it has grown to offer a wide variety of effective sampling algorithms to explore backbone, side-chain and sequence space. In this specific case I am working on coiled coils of three α helices supercoiled homo-dimeric bundles (fig .1) – homodimeric $3H5L_2A(H3H5L_2A)$, homodimeric $3H5L_2B(H3H5L_2B)$, homodimeric $3H5L_2C(H3H5L_2C)$. More in details the homo-dimeric structures are based on the antiparallel monomeric untwisted three-helix bundle five-layer ($3H5L_2$) with 80-residue helices and an 18-residue repeat unit (fig.2). heme binded The monomeric structure $3H5L_2$ was designed using parametric equations first derived by Francis Crick. This structure plays important roles in biology, and its simplicity and regularity have inspired peptide-design efforts. The parametric backbone generation was combined with the Rosetta protein-design

methodology to generate more complex and stable protein structures. The thermodynamic characterization showed that $3H5L_2$ was exceptionally stable with a denaturation midpoint of 7.5 M guanidinium chloride (GdmCl) at 25°C and 7M at 80°C. Moreover the 2.8 Å crystal structure of $3H5L_2$ has the same topology as that of the design model (D. Baker et al., Science 346, 481, 2014).

AlphaKnot: Server to analyze entanglement in structures predicted by AlphaFold methods

Wanda Niemyska, Pawel Rubach, Bartosz A. Gren, Mai Lan Nguyen, Wojciech Garstka, Fernando Bruno da Silva, Eric J. Rawdon, and Joanna I. Sulkowska

Interdisciplinary Laboratory of Biological Systems Modelling Centre of New Technology, Warsaw, Poland

We present AlphaKnot[1], the first server to assess entanglement of AlphaFold[2]-solved protein models with regard to its pLDDT data. Server has two main functionalities. One is a database of structures from all of 21 full proteomes solved by Alphafold which have been published up to 2022. Second is a user-friendly web server for researchers to analyze their own AlphaFold predictions. By using pLDDT confidence score, we classify predictions into categories which allows for detailed analysis, whether the protein model is correctly solved. This has allowed us to discover new types of knots in the human proteome[3]. By cross-validating AlphaFold predictions with our server and RoseTTa predictions, we show that AlphaFold, while overall a great tool, can have problems with correctly modeling knot topology of proteins. We show examples of AlphaFold models with wrongly predicted topology as well as give possible explanations of such occurrences.

References:

1. Niemyska, Wanda, et al. "AlphaKnot: Server to analyze entanglement in structures predicted by AlphaFold methods" Nucleic Acids Research [doi:10.1093/nar/gkn000].

2. Jumper, John, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596.7873: 583-589

3. Perlińska, Agata, et al. „New 6_3 knot and other knots in human proteome from AlphaFold predictions" [https://doi.org/10.1101/2021.12.30.474018]

EUROPEAN ROSETTACON
ON PROTEIN STRUCTURE
PREDICTION AND DESIGN

WARSAW
10-13
MAY
2022

SPONSORS:

Rosetta Commons

UNIVERSITY OF WARSAW