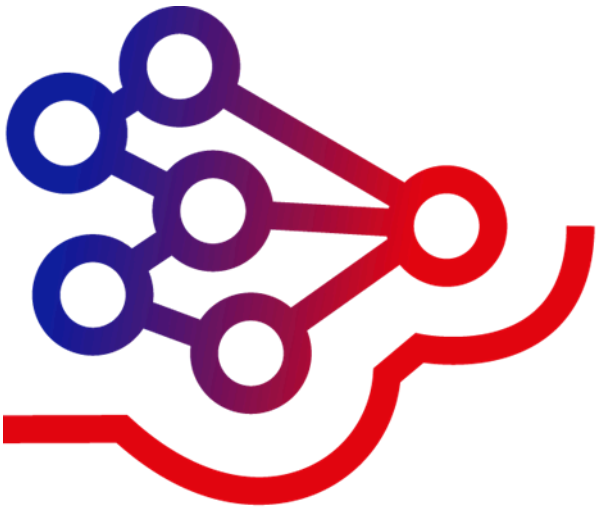

ML4NGP

MACHINE LEARNING FOR NON GLOBULAR PROTEINS

ABSTRACT BOOK

2nd MEETING on MACHINE LEARNING AND NON-GLOBULAR PROTEINS

14 – 17, May 2024

Hotel Onoma, Thessaloniki, Greece

cost
EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the European Union

KEYNOTE SPEAKERS



Prof. Christine ORENGO

University College London, UK



Prof. Michele VENDRUSCOLO

University of Cambridge, UK



Prof. Andrei N. LUPAS

MPI for Biology Tübingen, DE



Prof. Christos A. OUZOUNIS

Centre for Research and Technology Hellas, Greece

SELECTED TALKS

S01

The Cys sense: Regulation of cellular proliferation and protein homeostasis through redox-switch proteins

Meytal Radzinski[1], Tal Oppenheim[1], Rosi Fassler[1], Yair Rokach[1], Tommer Ravid[1], Merav Braitbard, [1,2] Dina Schneidman[2], Dana Reichmann[1]

[1]Department of Biological Chemistry, The Hebrew University of Jerusalem, Israel
[2]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel

Why do we age? How do we cope with environmental and physical changes? What makes pathogens smarter than their host? One of the biological processes linking all these questions relates to the amazing ability of cells to respond to changes in the environment, such as oxidative stress, and protect their proteome against damage. Here, I will cover different structural and functional mechanisms that redox-regulated proteins play in stress-response mechanisms, including maintaining protein homeostasis during oxidative stress and aging.
I will describe a toolbox of methodologies enabling us to discover novel, moonlighting redox-switch proteins that mediate diverse functions including cellular proliferation and protein quality control.
Specifically, I will focus on one of the key proteins in the protein degradation pathway, highly conserved AAA-ATPase, Cdc48 (VCP /p97), and its interaction with canonical co-factors. This study combined high-resolution cross-linking mass spectrometry (XL-MS) with integrative modeling to map the interaction between Cdc48 and its UBX co-factors and define its dynamics. We identified a potential redox-switch residue, cysteine 115, which is crucial for the stability of the ternary complex in vivo and in vitro. I will describe how the redox status of this cystine defines the activity of Cdc48 in cells as well as cellular survival during oxidative stress and aging. This study defined Cdc48's mode of interaction and can be used for the development of therapeutics mediating protein quality control.
To extend our knowledge regarding redox-switch proteins, we have established a robust methodology enabling us to track redox status at single-cell resolution using the in-vivo encoded redox-sensitive probes. This methodology was used to uncover the redox heterogeneity in cells and specific organelles, leading to the identification of distinct protein profiles and key redox switches defining the redox status of cells during aging, cell cycle, and cellular proliferation.

References:
Radzinski M., et al, eLife, 2018, 7: e37623
Reichmann et al., Mol Cell , 2018. 69(2):203-213
M. Radzinski and D. Reichmann, 2018, Current genetics, 65(1):301-306
Radzinski M., et al, Biomolecules, 2021, 11(3), 469
Oppenheim T, et al., Structure, 2023

S02

Combined use of NMR spectroscopy and computation to understand protein phase separation

Markus Zweckstetter

DZNE

"no information provided"

S03

Data-driven modeling of the intrinsically disordered proteome

Giulio Tesei, Anna Ida Trolle, Nicolas Jonsson, Johannes Betz, Frederik E., Francesco Pesce, Kristoffer E. Johansson, Kresten Lindorff-Larsen[1]

[1]University of Copenhagen, Denmark

Intrinsically disordered regions (IDRs) constitute about one third of the human proteome and play important roles in biological processes. While lacking well-defined 3D structures, IDRs adopt heterogeneous conformational ensembles which are determined by the amino acid sequence, yet not captured by existing methods for structure prediction. I will first present CALVADOS, an efficient coarse-grained molecular model of IDRs. The model was trained on experimental data reporting on conformational properties and validated extensively against both conformational and phase properties of IDRs with diverse sequences. Second, I will describe how we investigated relationships between sequence, conformational ensembles, and biological functions using molecular simulations of 28,058 human IDRs performed with CALVADOS. Through bioinformatics analyses, we explored the relationship between the conformational properties of the IDRs, which we obtained from simulations, and the biological function and cellular localisation of the full-length proteins. Further, we looked at how sequence features relate to conformational properties and analysed the evolutionary conservation of compaction across homologous IDRs. Our work exemplifies how a proteome-wide database of conformational ensembles can be used to gain insights into sequence–ensemble relationships and to formulate hypotheses on the biological role and evolution of human IDRs.

S04

Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 3

Alessio Del Conte, Mahta Mehdiabadi, Silvio CE Tosatto, Damiano Piovesan[1]

[1]University of Padova, Dept. of Biomedical Sciences, Padua, Italy

Protein intrinsic disorder (ID) is a complex and context-dependent phenomenon that encompasses a spectrum ranging from fully disordered states to folded states with long dynamic regions. The

absence of a universally applicable ground truth for all ID variations, coupled with the potential for order-to-disorder transitions under specific conditions, poses challenges in ID prediction.
The CAID challenge aimed to assess the performance of different prediction methods across diverse benchmarks, utilizing the annotation provided by the DisProt database. The DisProt database stores the coordinates of ID regions when there is experimental evidence in the literature.
In selecting a predictor, researchers may need to consider a trade-off between performance and execution time depending on the application. The CAID challenge revealed varying performance among different prediction methods across different benchmarks, emphasizing the ongoing need for the development of more versatile and efficient prediction software.
In this report, we present preliminary results from Round 3 of the CAID challenge. The performance of predictors was compared against baselines incorporating evolutionary information, disorder content, and a testing predictor utilizing AlphaFold2 pLDDT and solvent accessibility as proxies for disorder and binding inference.

S05

Blending Biology, Chemistry and AI to Enable Personalized Systems Pharmacology

Patrick Aloy[1]

[1]IRB Barcelona

Genomic initiatives are providing new target opportunities, but many correspond to gene products thought to be undruggable (i.e. non-globular proteins) and thus, unfortunately, these global approaches have not yet translated into new drugs. Big Data analytical techniques and AI have the potential to transform drug discovery, as they are reshaping other areas of science and technology, but we need to blend biology and chemistry in a format that is amenable for modern machine learning. In this talk, I will present the Chemical Checker (CC), a resource that provides processed, harmonized and integrated bioactivity data on small molecules. The CC divides data into five levels of increasing complexity, ranging from the chemical properties of compounds to their clinical outcomes. In between, it considers targets, off-targets, perturbed biological networks and several cell-based assays such as gene expression, growth inhibition and morphological profiles. I will also present the Bioteque, a resource of unprecedented size and scope that contains pre-calculated biomedical embeddings around 11 biological entities (e.g. genes, cells, tissues, disease, etc), derived from a gigantic knowledge graph, so that each entity can be described considering different contexts (e.g. interactions, expression, etc). With small molecule and biological bioactivity descriptors in hand, we now face a new scenario for chemical and biological entities where they both are translated into a common numerical format. In this computational framework, complex connections between entities can be unveiled by means of simple arithmetic operations. Indeed, we demonstrate and experimentally validate that these descriptors can be used to reverse and mimic biological signatures of disease models and genetic perturbations in vitro and in vivo, or downregulate the activity of SNAIL1, a non-globular transcription factor related to metastasis initiation thought undruggable.

References

Duran-Frigola et al. Extending the small molecule similarity principle to all levels of biology with the Chemical Checker. 2020. Nat Biotechnol. 38: 1087-1096.
Bertoni et al. Bioactivity descriptors for uncharacterized chemical compounds. 2021. Nat Commun. 12: 3932.
Pauls et al. Identification and drug-induced reversion of molecular signatures of Alzheimer's disease onset and progression in AppNL-G-F, AppNL-F, and 3xTg-AD mouse models. 2021. Genome Med. 13:168.
Fernández-Torras et al. Connecting chemistry and biology through molecular descriptors. 2022. Curr Opin Chem Biol, 66: 102090.
Fernández-Torras et al. Integrating and formatting biomedical data in the Bioteque, a comprehensive repository of pre-calculated knowledge graph embeddings. 2022. Nat Commun. 13: 5304.

SHORT TALKS

T01

Structural studies of mineral interactions with intrinsically disordered magnetosome-associated proteins

Raz Zarivach

The process of biomineralization, which involves the formation and control of minerals, relies on specialized proteins. These proteins often have regions that lack a specific structure but play a crucial role in their biomineralization function. To unravel this intricate process, our investigation focused on ferritins and disordered protein segments engaged in mineral interactions, specifically within magnetotactic bacteria. We unveiled the dynamic interplay between magnetite-interacting components within magnetosome-associated proteins and iron oxide particles by employing a comprehensive suite of advanced techniques, including NMR spectroscopy, X-ray crystallography, and cryo-electron microscopy. Using this approach, we showed the direct interaction of minerals and proteins for the first time. Notably, we established a novel connection between the folding dynamics of Intrinsically Disordered Protein (IDP) segments and their recognition and interaction with mineral entities. These findings spotlight the innate capability of ferritins and disordered proteins to undergo conformational adaptations in response to environmental cues, akin to versatile performers adjusting to a changing stage. These findings provide valuable insights into the role of disordered proteins in controlling the structure of minerals during biomineralization processes.

T02

Spider silk: from NMR structural studies to mechanism of formation and high-performance artificial fibres

Kristaps Jaudzems[1]

[1]Latvian Institute of Organic Synthesis

Spider silk is one of nature's most amazing biomaterials. It is composed of proteins called spidroins that aggregate into fibers with a β-sheet secondary structure. Understanding and reproducing the exact mechanism of silk formation is crucial for obtaining artificial spider silk fibers with native-like properties, which could have applications in regenerative medicine, drug delivery, and other fields.
The formation of spider silk occurs in the spider's silk gland, where changes in environmental conditions facilitate the structural conversion of the spidroins into solid fibers. We employ nuclear magnetic resonance (NMR) to simultaneously assess protein structure and stability at the residue or even atomic level. Thus, we have mimicked the process of spider silk formation in the NMR tube and studied the role of individual silk protein domains [1]. Based on our findings, we have proposed a stepwise mechanism of structural changes required to obtain high-performance fibers in a biomimetic way.
Finally, we use magic angle spinning solid-state NMR to characterize the structure of artificial spider silk fibers obtained using a biomimetic approach. These experiments provide valuable information

on how the protein structure is affected by the silk protein sequence, spinning technique, and post-spinning treatments [2].

[1] N. Kronqvist et al. Nat. Commun., 5, 3254 (2014); M. Andersson et al. PLoS Biol., 12, e1001921 (2014). M. Otikovs et al. ChemBioChem, 16, 1720–1724 (2015); M. Otikovs et al. Angew. Chem. Int. Ed., 56, 12571–12575 (2017); M. Sarr et al. J. Biol. Chem., 101913 (2022); M. Šede et al. Front. Mol. Biosci., 9, 936887 (2022).
[2] M. Otikovs et al. Angew. Chem. Int. Ed., 56, 12571–12575 (2017); P. Mohammadi et al. Sci. Adv., 5, eaaw2541 (2019); T. Arndt et al. Adv. Funct. Mater., 2200986 (2022); T. Arndt et al. Nat. Commun., 13, 4695 (2022).

T03

A disordered loop in a bacterial adhesin can form a poly-proline II helix upon binding to heparin

Ina Meuskens, Per Eugen Kristiansen, Benjamin Bardiaux, Nadia Izadi-Pruneyre, Dirk Linke

Background: many bacterial adhesins in pathogenic species interact with the extracellular matrix of host cells to initiate contact. This triggers further processes that ultimately can lead to cell or tissue invasion. It is usually assumed that a specific adhesin from a given bacterial species will bind to a specific receptor or matrix molecule.
Objectives: our objective is to better understand the relevance of strain-specific variations in bacterial adhesins, and how these might contribute to different clinical outcomes between different isolates of the same species.
Methods: to better understand the binding interactions between our adhesin variants and the extracellular matrix, we use biophysical methods such as surface plasmon resonance, small angle x-ray scattering, ELISA-based binding assays, NMR, and computational modelling. Most of our assays use both purified adhesins, or whole bacterial cells that recombinantly express the adhesin in question.
Results: a short insertion of 30 amino acids in the head domain of the bacterial adhesin YadA from enteropathogenic Yersiniae changes the binding mode of the adhesin, switching from collagen binding to binding negatively charged glycans, including heparin. The insertion forms a polyproline II helix upon binding, shown by NMR spectrosopy. I will discuss the physiological and evolutionary relevance of the binding partners, and how the strain-specific difference in binding might lead to different clinical outcomes.

Reference: Meuskens et al., A poly-proline II helix in YadA from Yersinia enterocolitica serotype O:9 facilitates heparin binding through electrostatic interactions. FEBS Journal, in press

T04

THE ROLE OF ENVIRONMENT IN AMYLOID AGGREGATION

Vytautas Smirnovas

The ability to form amyloid structures may be a generic property of polypeptides, and there are two major factors which define the probability of amyloid fibril formation – amino acid sequence of the protein/peptide and the environmental conditions. In the case of folded proteins, at least partial unfolding is necessary to trigger the amyloid formation pathway, so increased temperature, extreme pH conditions, addition of denaturants or any other changes in the environment leading to destabilization of protein structure are used in amyloid aggregation studies. Even in the case of disordered proteins, neutralization of charges or contact with hydrophobic surfaces may be necessary to induce amyloid formation. In addition to the specific conditions required for amyloid formation, changes in the environment may alter the mechanism of aggregation and lead to distinct amyloid fibril conformations. Finally, environment conditions affect the kinetics of aggregation and may alter the effect of anti-amyloid compounds.

T05

Cryo-EM structures of functional and pathological amyloid ribonucleoprotein assemblies

Javier Garcia-Pardo, Andrea Bartolomé-Nafría, Salvador Ventura[1]

[1]Universitat Autònoma de Barcelona, Spain

Amyloids are implicated in neurodegenerative and systemic diseases, yet they serve important functional roles in numerous organisms. Heterogeneous nuclear ribonucleoproteins (hnRNPs) represent a large family of RNA-binding proteins (RBPs) that control central events of RNA biogenesis in normal and diseased cellular conditions. Many of these proteins contain prion-like sequences of low complexity, which not only assemble into functional fibrils in response to cellular cues but can also lead to disease when missense mutations arise in their sequences. Recent advances in cryo-electron microscopy (cryo-EM) have provided unprecedented high-resolution structural insights into diverse amyloid assemblies formed by hnRNPs and structurally related RBPs, including TAR DNA-binding protein 43 (TDP-43), Fused in Sarcoma (FUS), Orb2, hnRNPA1, hnRNPA2, and hnRNPDL-2. This review provides a comprehensive overview of these structures and explores their functional and pathological implications.

T06

Hidden flexible state(s) of sigma-1.1 domain of RNA polymerase

Dávid Tužinčin[1,2], Petr Padrta[2], Hana Šanderová[3], Alžběta Rabatinová[3], Kateřina Bendová[1,2], Libor Krásný[3], Lukáš Žídek[1,2], Pavel Kadeřávek[1,2]

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

[3]Laboratory of Microbial Genetics and GeneExpression, Institute of Microbiology, CzechAcademy of Sciences, Vídeňská 1083, 142 20, Prague, Czech Republic

It is well known that determining the low-energy structure of a protein often does not fully elucidate its purpose and behavior in living systems. Proteins can undergo various conformational changes crucial for their function. Here, we investigated the sigma-1.1 domain of vegetative sigma factors of bacterial RNA polymerase. While it has been previously described as a globular protein, we found that such a description is insufficient to depict its role in bacteria comprehensively. The major state of the protein is indeed a compact structure in agreement with the previous work, but it is not its sole form and we detected a second state in exchange with the major state. The minor state comprises only about 4% of the population at room temperature. Such a sparsely populated state can be easily overlooked by standard methods used to study protein structures. To investigate this minor state, we employed specialized NMR methods. Our findings revealed that the transition to the minor state affects the hydrophobic core of the protein and is accompanied by a significant decrease in secondary structure propensities. This suggests that the globular form of the protein is no longer preserved, and the protein becomes significantly more disordered and flexible. Studying the temperature dependence of the population of the minor state, along with investigating the thermal stability of the protein, showed that the proportion of the less ordered state(s) increases with temperature, becoming quite significant at higher temperatures. Functional characterization of the sigma-1.1 domain provided evidence of the biological relevance of the detected more disordered state(s) for RNA polymerase function and it was demonstrated that increased plasticity is crucial for accelerating transcription at elevated temperatures.

T07

Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins

Javier González-Delgado[1], Pau Bernadó[2], Pierre Neuvial[3], Juan Cortés[1]

[1]LAAS-CNRS, Institut de Mathématiques de Toulouse, Université de Toulouse
[2]Centre de Biologie Structurale, Université de Montpellier
[3]Institut de Mathématiques de Toulouse, Université de Toulouse

The high conformational variability of flexible proteins makes their structural characterization a non-trivial task. The approaches commonly used to describe proteins on the basis of a single or small number of structures are unsuitable in this context. Although some extensions have been proposed using average-based techniques, their practical applicability remains limited to proteins that fluctuate around a stable conformation, but they are not capable of disentangling the underlying variability of highly-flexible system such as intrinsically disordered proteins. This work proposes to extend the concept of contact maps to the ensemble framework by incorporating the

intrinsic probabilistic nature of disordered proteins. In this regard, a conformational ensemble is characterized through a weighted family of contact maps. To do so, conformations are first described with a refined definition of contact that suitably accounts for the geometry of the inter-residue interactions and the sequence context. Then, a clustering algorithm is performed to retrieve representative contact patterns. The performance of the method is illustrated by several use-cases and compared with other existing approaches that prove ineffective in capturing relevant structural features of highly-flexible proteins.

T08

Development of machine learning algorithms for the prediction of the binding affinity between peptides and their target proteins

Che Yee Ng[1], Gerry A.F. Nicolaes[1,2], Kanin Wichapong[1,2]

[1]Department of Biochemistry, Cardiovascular Research Institute Maastricht (CARIM), Maastricht University, Maastricht, the Netherlands
[2]Hillmark B.V, Maastricht the Netherlands

Peptide-based drugs offer the potential to disrupt disease-related protein-protein interactions. However, development of therapeutic peptides poses challenges, particularly in predicting binding affinity (BA) with protein targets. While computational tools like molecular dynamics (MD) simulations and binding free energy (BFE) calculations are widely used in drug discovery, they are computationally demanding. Despite machine learning (ML) algorithms showing a good performance in predicting BA for small-molecule inhibitors, their performance with peptides has yielded poor results. In this study, we aim to integrate MD simulations and ML to create models enabling rapid and accurate prediction of BA between peptides and protein targets.
A curated dataset containing protein-peptide complexes (n = 186) was obtained from the PDBbind database (v.2020). Each complex was subjected to a short MD simulation (3 ns), repeated five times (replica = 5). Subsequently, BFE between proteins and peptides was calculated to generate essential descriptors including electrostatic (EELE), Van der Waals (EVdW) interactions and free energy of solvation (Gsolv). Initially, a linear regression model based on these 3 features was generated and the derived model exhibited a low correlation between predicted and experimental BA values after exclusion of the non-significant models (R = 0.40 and R2 = 0.12, averaged over 17 runs after exclusion of 3 models). Thus, six different ML algorithms were explored to evaluate their predictive efficacy. Employing the CatBoost algorithm yielded an enhanced R-value of 0.44, and integration of one additional molecular descriptor (peptide_amino_acid_length) slightly increased the R-value to 0.49 (both averaged over 19 runs as 1 run was excluded). Notably, data partitioning using random or stratified methods showed negligible influence on outcomes (RCatBoost_random_split = 0.49 vs. RCatBoost_stratified_split = 0.47, exclusion of 1 and 3 models respectively). In conclusion, the preliminary results suggest that integrating relevant peptide descriptors alongside structural MD features could boost ML model performance. To improve the model, we will therefore include several molecular features e.g. the number of rotatable bonds, total charges, and those will be combined with different MD parameters like various free energy of solvation models. Ultimately, the developed

model should offer a rapid and accurate means to predict the BA of peptides across diverse protein targets.

T09

Pioneer factor Sox2 remodels nucleosomes and displaces histone H1

Sveinn Bjarnason[1], Jens Nicolai Vilstrup Decker[1], Eliška Koutná[2], Kinga S. Demény[1], Vaclav Veverka[2], Pétur O. Heidarsson[1]

[1]Department of Biochemistry, Science Institute, University of Iceland
[2]Department of Cell Biology, Faculty of Science, Charles University

Through their interaction with chromatin, pioneer transcription factors not only initiate gene expression changes essential for development but also serve as key regulators in cellular reprogramming and disease processes. The process by which the pioneer transcription factor Sox2 remodels nucleosomes and its capacity to displace histone H1 remains poorly understood. Similarly, the structural configuration of Sox2's intrinsically disordered regions (IDRs) when engaged with nucleosomes, and the subsequent effects on nucleosome conformation, are not well defined. In an attempt to address these gaps, we employed single molecule Förster Resonance Energy Transfer (smFRET) and Nuclear Magnetic Resonance (NMR) spectroscopy to elucidate the conformations of Sox2 in complex with nucleosomes. We explore its binding modes, the location dependent affinity of binding sites and the resulting structural impacts on the nucleosome itself. Our findings reveal that Sox2's IDR extends significantly upon interaction with the nucleosome, and we identify a specific interaction site within the IDR. Interestingly, Sox2's binding affinity remains consistent, irrespective of the location of the binding site. Moreover, our data not only confirms Sox2's capability to displace linker histone H1, highlighting a critical mechanism in chromatin remodeling, but also positions this activity within the wider context of pioneer transcription factors strategies for chromatin engagement. Identifying an interaction site within Sox2's IDRs marks a step forward in our grasp of pioneer transcription factor mechanism, offering promising leads for development strategies aimed at fine-tuning Sox2's function in cellular reprogramming and oncogenesis.

T10

Residue Interaction Network Generator v4: improved interaction recognition for all 35,000 PDB chemical components

Alessio Del Conte[1], Giorgia Francesca Camagni[1], Damiano Clementel[1], Alexander Miguel Monzon[2], Giovanni Minervini[1], Damiano Piovesan[1], Silvio Tosatto[1]

[1]University of Padua, Department of Biomedical Sciences
[2]University of Padua, Department of Information Engineering

The Residue Interaction Network Generator (RING) is a powerful tool designed to identify non-covalent interactions within protein structures at the atomic level. It efficiently processes multi-state structures, including molecular dynamics and structural ensembles, to generate probabilistic networks and conformation-dependent contact maps. Notably, RING exhibits remarkable speed and scalability, with processing times scaling very well with input size.
In its latest iteration, RING has significantly expanded its capabilities by integrating the structures of all chemical components found in the Protein Data Bank (PDB). This integration includes the definition of hydrogen atom donors and acceptors, accurate addition of hydrogen atoms with correct nomenclature and placement, and the generalization of the detection of aromatic rings. Moreover, this new version has expanded its set of interactions to include novel ones such as Pi-hydrogen, metal ion coordination, halogen interactions, and updating the former ones with new state-of-the-art definitions. With support for over 35,000 chemical components, RING empowers users to explore diverse molecular interactions within their protein structures comprehensively.
The RING ecosystem also encompasses the RING web server, a user-friendly interface that extends the functionality of the software to a broader audience. Through the web server, users can conveniently access RING's capabilities without the need for installation and utilize computational resources. The web server now allows users to visualize structures with added hydrogens, explore various interactions calculated by RING with detailed information directly on the 3D structure, paired with many other components for different visualizations of the contacts.
Overall, the latest version of RING represents a significant leap forward in the analysis of protein structures, offering enhanced functionality, expanded molecular interaction capabilities, and improved usability for researchers in the field of structural biology and drug discovery either directly or through the use of an easy-to-use web server.

T11

FAIR Research Software as the catalyst for trustworthy AI in Life Science

Fotis Psomopoulos[1]

[1]CERTH, Greece

Ensuring that data are FAIR is nowadays a clear expectation across all science domains, as a result of many years of global efforts. Research software, has only just started to receive the same level of attention in recent years, with targeted actions towards the definition of the FAIR principles as applied to research software, as well as concerted efforts around reproducibility, quality, and sustainability. Given the rapid rise of ML as a key technology across all science domains, it is important to build on our collective experience, and already start addressing the challenges ahead of us, towards making ML FAIR.

T12

Understanding specific interactions in phase-separated Germ granules by multi-scale simulations

Vasileios Xenidis[1], Arya Changiarath Sivadasan[2], Kumar Gaurav[1], Spyridon Papakostas[3], Lukas Stelzl[2]

[1]Aristotle University of Thessaloniki, Greece
[2]Johannes Gutenberg University of Mainz, Germany
[3]International Hellenic University

Deciphering the influence of protein sequence on condensate formation poses a significant challenge in theoretical chemistry. While molecular dynamics simulations coupled with free-energy calculations provide valuable insights into the chemical and biological properties of protein sequences, their computational expense prompts the emergence of more efficient methods, such as coarse-grained (CG) simulations. In the context of key intrinsically disordered regions (IDRs) H-I, J, K, and L within MUT-16, recognized as pivotal for mutator focus formation, CG simulations targeting the region spanning amino acids 633-944 unveil critical factors like protein size and the abundance of aromatic amino acids. These simulations demonstrated substantial condensate formation, aligning with literature and underscoring the region's role in promoting protein aggregation and phase separation. The condensate sequence exclusively features phenylalanine and tyrosine, impactful amino acids known to influence phase separation behavior. Their abundance or absence significantly influences condensate size, emphasizing the intricate relationship between protein size, aromatic amino acid content, and phase separation efficiency in MUT-16. Thereafter, in order to further study sequence-property relationships, a state-of-the-art Protein Language model was employed. These models not only carry the potential for reducing the required training data through active learning but also have the potential to aid in the understanding of the binding mechanism between MUT-16 condensate and disordered peptides. This model generated, before and after fine-tuning, intrinsically disordered peptides (51-amino-acid long). The fine-tuning process involved training with IDPs from the DisProt database. Afterwards, disorder scores in both cases were assessed through deep learning. The results showed a notable increase in disordered peptide production following LLM fine-tuning. Moreover, the observed amino acid frequencies revealed nuanced yet impactful variations. Despite being a prominent promoter of disordering, amino acid P is employed slightly less frequently after fine-tuning, highlighting the critical influence of sequencing and amino acid identity on disordering impact. In this context, the focus is on examining whether specific key amino acid motifs, akin to 'words' in the language of amino acids, are present. Additionally, the investigation extends to broader sequence patterns, analogous to the 'grammar' that influences the disordering process and subsequent phase separation. Ongoing research integrates CG simulations, free-energy calculations, and a different LLM employing Bayesian Optimization. We envisage that this methodology will help us to better understand sequence-property (condensate binding) relationships.

T13

Local Energetic Frustration and Intrinsic Protein Disorder

Maria I Freiberger[1], Alexander M. Monzon[2], R. Gonzalo Parra[3]

[1]Buenos Aires University, Argentina
[2]Padua University, Italy
[3]Barcelona Supercomputing Center, Spain

Foldable proteins satisfy the "minimum frustration principle" meaning that energetic conflicts between different residues are minimized in their native states. Because of this, the energy landscape of natural and foldable proteins resembles a funnel with a strong bias towards the native state located in a deep energetic minimum at its bottom. However, not all interactions that are present in the native state of proteins can be energetically minimized and some remain in energetic conflict, i.e. they are highly frustrated. These conflictive signals have been shown to be associated with several functional aspects of proteins such as protein-protein binding sites, allosterism, and catalysis. More recently, we have shown that functional frustration is conserved across members of the same protein family, underscoring its evolutionary importance.
Intrinsically disordered proteins (IDPs) represent a set of molecules that, in the most extreme case, cannot adopt a defined tertiary structure on their own and instead exist as ensembles of (non-random) unfolded, non-globular conformations. It is believed that the energy landscapes of IDPs are more shallow than the ones of foldable proteins due to the presence of extensive energetic conflicts within the polypeptide chain, i.e. they are more highly frustrated. Nuclear Magnetic Resonance (NMR) constitutes a technique that allows us to analyze the less extreme case of highly flexible regions within partially disordered proteins.
To study the relationship between local energetic frustration and intrinsic protein disorder we have analyzed local frustration patterns across all experimental NMR ensembles that are present in the Protein Data Bank. We have used our FrustratometeR algorithm to calculate local frustration across all the conformations within our NMR dataset and IUPRED to define ordered and disordered residues. We have analyzed frustration and disorder in the context of solvent accessibility and secondary structure compositions. We have characterized the tendencies for all these properties as a function of amino acid identities.
Our results constitute the first extensive analysis where an explicit link is established between local frustration and intrinsic disorder in proteins. We offer an energetic classification of amino acids and their propensity to introduce protein disorder in foldable proteins.

T14

Disorder of disease-related proteins - a case study of IGDD

Lazar Vasović, Jovana Kovačević[1]

[1]Faculty of Mathematics, Belgrade University

Several publicly available databases offer information regarding the connection between genes and diseases in different formats. In order to facilitate the usage of these databases, we integrated several of them in a uniform database called Integrated Gene Disease Database (IGDD). In IGDD, we included gene-disease connections from the following databases: DisGeNet, Cosmic, HumsaVar, Orphanet, ClinVar, HPO и Diseases. IGDD contains around 300 000 connections between genes and diseases, including additional information: Gene Symbol, Entrez ID, UniProt ID and Ensembl ID, regarding the gene and the corresponding protein, DOID and Disease Name regarding the disease and Source, representing the database where the gene-disease connection was taken from. In order to enrich IGDD, information about the disorder of the gene-corresponding proteins was added. The disorder was determined computationally using a few disorder predictors, including Iupred and VSL2 for each protein. IGDD file unites all the gene-disease connections from different databases in a unique format with additional information and thus provides a valuable resource for future data analysis and further investigation.
In this research, we carried out a quantitative analysis of IGDD with the aim of checking whether there is a relationship between certain diseases and the level of disorder of proteins connected to them. Additionally, using disease ontology (DO) identifiers, we examined whether the diseases with the same ancestor in the DO share similar patterns in level of disorder of the connecting proteins.

T15

Transferable deep generative modeling of intrinsically disordered protein structural ensembles

Giacomo Janson, Michael Feig[1]

[1]Department of Biochemistry and Molecular Biology, Michigan State University, USA

MOTIVATION: The characterization of structural ensembles of intrinsically disordered regions (IDRs) of proteins is crucial for obtaining a deeper understanding of their function. Studying such ensembles is a challenging problem requiring the integration of experimental and in silico approaches. Simulation methods, like molecular dynamics, are valuable tools for constructing IDR ensembles, albeit with high computational costs. Recently, machine learning (ML) models that learn from simulations have shown promise as an efficient alternative for generating ensembles. Yet, ML methods currently have limited transferability when applied to proteins outside the training data. Here, we describe a novel deep generative model that demonstrates high levels of transferability in IDR modeling.
METHODS: Our approach, named idpSAM (Structural Autoencoder Model), is a latent diffusion model. Diffusion models are a powerful type of generative models that have state-of-the-art performance in protein structure generation tasks. IdpSAM combines an autoencoder to learn a representation of protein geometry and a diffusion model to sample conformations in the encoded space. This design choice enables idpSAM to couple computational efficiency with the capacity to model complex data. The model currently operates only on Cα atoms, but atomistic details can be recovered via the cg2all network developed by our group. We trained idpSAM on a large dataset of atomistic simulations of >3,000 IDRs from DisProt, with lengths ranging from 15 to 60 residues. The simulations were performed with the ABSINTH implicit solvent model, a method providing a good tradeoff between computational efficiency and the ability to capture complex structural features, such as transient secondary structure elements.

RESULTS: IdpSAM faithfully reproduces structural ensembles of test sequences absent in the training set. Notably, we demonstrate how the model can capture even subtle differences between ensembles of wild type and mutant IDRs differing by a few amino acids. We also investigate the important relationship between the model's transferability and its training set. We show how the performance of a ML ensemble generator varies depending on the size and diversity of the training set. Finally, we illustrate as a proof of principle how external data in the form of distance restraints can be incorporated to bias the ensemble generation process.
CONCLUSIONS: We believe that idpSAM represents a significant progress in IDR ensemble modeling via ML and that it can potentially be adapted to more complex data and integrate experimental restraints.

T16

Developing machine-learning-based amyloid predictors with Cross-Beta DB.

Valentin Gonay[1,2], Michael P. Dunne[2], Javier Caceres-Delpiano[3], Andrey V. Kajava[1]

[1]CRBM, UMR 5237 CNRS, Université Montpellier, Montpellier, France
[2]PROTERA SAS, 176 avenue Charles de Gaulle, 92522 Neuilly-sur-Seine Cedex, France
[3]Protera (GEAEnzymes SpA), Av. Santa Maria 2810, Oficina 302 Providencia, Santiago, Chile

Numerous studies have demonstrated that the propensity to form amyloids is coded by the amino acid sequence and this finding has paved the way for the development of several computational predictors of amyloidogenicity. The ultimate objective of computational methods is to accurately predict the formation of disease-related and functionally relevant amyloids that occur in vivo. These amyloid fibrils are known to form very specific ""cross-β"" structures of protein regions longer than about 15 residues. Remarkably, despite the significance of the naturally occurring amyloids, there has been a lack of datasets specifically dedicated to them. Hence, we built Cross-Beta DB, a database composed of cross-β amyloids formed in natural conditions. This database is expected to be indispensable for benchmarking amyloid predictors. We used the Cross-Beta DB to train and benchmark several such algorithms, using machine learning. The best-performing of these, the random-forest-based Cross-Beta RF Predictor, demonstrated superior performance over the other existing methods, fostering high expectations for an improved prediction of naturally occurring amyloids.

Keywords: amyloid, protein aggregation, database, machine learning, bioinformatics, cross-β structure, classification, benchmark
https://doi.org/10.1101/2024.02.12.579644

T17

Automatic Classification of Units in Tandem Repeat Proteins

Michela Quadrini, Luca Mariani, Diego Pennesi, Filippo Reucci, Luca Tesei[1]

[1]School of Science and Tecnology, University of Camerino, Via Madonna delle Carceri, 8, 62032, Camerino, Italy

Tandem repeat proteins (TRPs), a widespread class of non-globular proteins, are involved in many biological processes, playing essential roles in cell adhesion, protein-protein interactions, and molecular recognition. A TRP consists of Repeat Regions, formed by Insertions and Repeat Units, the smallest structural building block forming the repeat region. The unit classification plays a fundamental role in different tasks, including phylogenetic studies and understanding sequence-structure-function relations. Different and manually-based classifications have been proposed in the literature. Nevertheless, automatic classifications are of great value due to the increasing number of TR protein structures.
Method and Results.
In the talk, we introduce an automatic method to classify units in protein tandem repeats. The method abstracts the units by non-covalent interactions at the atomic level [1] and exploits Algebraic Structural Alignment (ASA) distance to quantify the structural dissimilarity between units [2]. The protein abstraction consists of five chemical interactions extracted by exploiting geometrical parameters. Such an abstraction is formalized as arc-annotated sequences (AASs), a base sequence equipped with arcs representing the chemical interactions. The structural differences between two units are quantified by ASA distance, a computational dissimilarity measure, which maps the AASs into structural trees that formalize arcs and relations among them by exploiting classical tree alignment. The hierarchical Clustering algorithm infers an automatic classification by grouping similar objects into clusters.
The pipeline related to the method is implemented in Python by exploring RING and STAlign, two tools to extract protein interactions and qualify the structural difference based on ASA distance, respectively. Moreover, the pipeline uses the implementation of the Hierarchical Clustering method provided in scikit-learn [3].
The approach has been tested on subsets of molecules of REPEAST DB [4], a database of annotated TRPs, and compared with the TM-ALIGN tool in order to classify the Class and Topology. The results, evaluated with the clustering methods, show that our approach overcomes the results obtained with the TM-ALIGN tool.
Conclusions.
This study introduces an automatic annotation method for TRP units that can be extended to other bioinformatics applications. We are working on the method extension by considering other protein abstractions and machine learning approaches combining sequences and structures.

References
[1] Clementel, Damiano, et al. ""RING 3.0: Fast generation of probabilistic residue interaction networks from structural ensembles."" Nucleic Acids Research 50.W1 (2022): W651-W656
[2] https://github.com/bdslab/STAlign
[3] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
[4] Paladin, Lisanna, et al. ""RepeatsDB in 2021: improved data and extended classification for protein tandem repeat structures."" Nucleic Acids Research 49.D1 (2021): D452-D457.
This study received funding from the European Union - Next-GenerationEU - National Recovery and Resilience Plan (NRRP) – MISSION 4 COMPONENT 2, INVESTMENT N. 1.1, CALL PRIN 2022 PNRR D.D. 1409 of 14th Sep 2022 - RNA2FUN CUP N. J53D23014960001."

T18

LCR-dataset: manually curated benchmark for training and fine-tuning language models to retrieve information on Low Complexity Regions and its functional annotations from scientific literature.

Sylwia Szymanska, Aleksandra Gruca[1]

[1]Department of Computer Networks and Systems, Silesian University of Technology, Gliwice, Poland

Low Complexity Regions (LCRs) are characterized by a limited set of amino acids in a protein sequence. These regions show important properties, often playing critical roles in protein-protein interactions, molecular recognition and signaling pathways.
One of the main sources of our knowledge of LCRs function is scientific literature, however the number of articles on this topic is scarce and finding them requires skills in search techniques. On the other hand, machine learning approaches, especially methods based on language models, can help researchers retrieve papers of interest.
In this work we introduce a novel, manually curated dataset, which is a collection of publications describing functions of LCRs. The LCR-dataset can be used in two ways: (1) as a comprehensive resource of papers describing particular LCR functions (2) as a benchmark dataset for training and fine-tuning machine learning models for the task of scientific publication retrieval.
Our dataset contains titles and abstracts from manually curated articles describing seven LCRs functions: antigen recognition (6432 papers), helix (1128), kinase activity (1935), nucleic acid binding (1348), phosphatase activity (311), transcription factor (1487) and zinc finger (407). Additionally we enrich our dataset with a collection of papers that include names similar to LCRs. Distinguishing similar tokens in token-level embeddings is difficult due to their close vector values, which makes classification more challenging. Finally, we define seven classification tasks with different difficulty levels starting with a simple binary classification and ending with a multi-classification problem.
To ensure reproducibility in the classification tasks, the LCR-dataset is divided into train, test, finetune and validation datasets. To generate embeddings, we used transformer-based architecture models such as BERT and BioBERT at the token-level, and SPECTER at the document-level. To handle the dimensionality reduction problem in token-level embeddings, we applied PCA to get one representative vector for each publication.
We provide baseline results for each task by evaluating fine-tuned transformer models with the Random Forest classifier. In a binary classification task, we successfully identify publications describing LCRs and their associated functions. Furthermore, we extend our analysis to multiclass classification, where classes are defined based on LCR and specific function in text. In this case we also used multiclass data to pretrained models and compare to baseline models. Providing classification tasks of varying complexity makes our dataset an interesting resource that can help to develop machine learning approaches which can be used to extend and systematize our knowledge on LCRs and their functions.

T19

Deciphering Allosteric Pathways

Turkan Haliloglu

Allostery, a fundamental biological control mechanism, is investigated here through an implementation of Gaussian Network Model (GNM)-based Transfer Entropy (TE) approach. This method dissects dynamic information into distinct subsets of slow modes, revealing e multi-directional allosteric pathways within protein structures. Within these subsets, the degree of collectivity in information transfer among residues emerges as a key dynamic characteristic, serving as global information hubs that drive functional motion within proteins. Notably, this dynamic phenomenon reflects causal relationships imprinted in the protein topology and yet persists with only sequence information, suggesting broader applicability. Across diverse protein structures, including non-globular ones, GNM-TE offers a promising methodology for identifying functional sites of collective influence and elucidating allosteric communication pathways.

T20

Optimizing energy functions with neural networks

Gábor Erdős, Zsuzsanna Dosztányi[1]

[1]MTA-ELTE Momentum Bioinformatic Research Group, Hungary

The investigation into protein dynamics remains fundamental in both basic research and industrial applications. Recent advancements in neural network technologies have revolutionized protein structure prediction, enabling the generation of structures suitable for molecular dynamic simulations. These simulations provide valuable insights crucial for drug development. However, intrinsically disordered proteins pose a unique challenge, as existing energy functions fail to adequately describe their dynamics, particularly due to the complex interplay of hydrophobicity.
In this study, we present a novel neural network architecture designed to optimize complex statistical potential-based force fields. Inspired by the work of PD Thomas and Ken Dill (1996), who developed a force field based on maximizing the likelihood of natural protein structures compared to artificial ones generated via gapless threading. We translated the problem into a network architecture that enables rapid and accurate optimization of the forcefield. Leveraging feedforward neural networks, our method exhibits exceptional scalability, enabling optimization on significantly larger datasets for more precise potential calculations.
Our findings demonstrate that even simple contact potential-based force fields can effectively describe globular structures. Moreover, we propose a transfer learning approach that allows for direct prediction of potential energy from protein sequences, facilitating predictions related to structure, such as disorder propensity, identification of conditional or functional disordered segments, and ensemble generation.

T21

aSynPEP-DB: a database of biogenic peptides for inhibiting α-synuclein aggregation

Carlos Pintado-Grima[1], Oriol Bárcenas[1], Valentín Iglesias[1], Jaime Santos[1,2], Zoe Manglano-Artuñedo[1], Irantzu Pallarès[1], Michał Burdukiewicz[1,3], Salvador Ventura[1]

[1]Universitat Autònoma de Barcelona, Spain
[2]Heidelberg University, Germany
[3]Medical University of Białystok, Poland

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, yet effective treatments able to stop or delay disease progression remain elusive. The aggregation of a presynaptic protein, α-synuclein (aSyn), is the primary neurological hallmark of PD and, thus, a promising target for therapeutic intervention. However, the lack of consensus on the molecular properties required to specifically bind the toxic species formed during aSyn aggregation has hindered the development of therapeutic molecules. Recently, we defined and experimentally validated a peptide architecture that demonstrated high affinity and selectivity in binding to aSyn toxic oligomers and fibrils, effectively preventing aSyn pathogenic aggregation. Human peptides with such properties may have neuroprotective activities and hold a huge therapeutic interest. Driven by this idea, here, we developed a discriminative algorithm for the screening of human endogenous neuropeptides, antimicrobial peptides and diet-derived bioactive peptides with the potential to inhibit aSyn aggregation. We identified over 100 unique biogenic peptide candidates and ensembled a comprehensive database (aSynPEP-DB, https://asynpepdb.ppmclab.com/) that collects their physicochemical features, source datasets, AlphaFold-predicted structures and additional therapeutic-relevant information, including their sites of expression and associated pathways. Besides, we provide access to the discriminative algorithm to extend its application to the screening of artificial peptides or new peptide datasets. aSynPEP-DB is a unique repository of peptides with the potential to modulate aSyn aggregation, serving as a platform for the identification of previously unexplored therapeutic agents or new biomarkers to monitor in PD patients.

T22

From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2

Hélène Bret, Jinmei Gao, Diego Javier Zea, Jessica Andreani, Raphaël Guerois[1]

[1]Université Paris-Saclay, CEA, CNRS, I2BC, France

The transformative power of AlphaFold2 has unlocked new ways for modeling and understanding protein-protein interaction networks, particularly those involving intrinsically disordered regions (IDRs). These regions, characterized by their flexibility and lack of a stable structure, are pivotal in mediating interactions within the cell. Nevertheless, their transient nature complicates the identification of their interaction interaction modes. Our study leverages the capabilities of

AlphaFold2-Multimer to tackle this challenge by employing a novel approach that integrates protein fragmentation and different strategies to build the input Multiple Sequence Alignments (MSAs).
Utilizing a dataset of protein-peptide complexes not redundant to the structures used by AlphaFold2 during training, we explored to delineate the interaction sites more precisely. Our findings reveal that the traditional full-length sequence approach yields a modest success rate of 40% in identifying the correct interaction sites. However, by fragmenting these regions into smaller, more manageable segments and employing a variety of MSA combination techniques, we were able to boost this success rate significantly to 90%.
We have validated the efficacy of our fragmentation and MSA-building strategies through their application to a broader dataset derived from the Eukaryotic Linear Motif (ELM) database. This extensive validation underscores the robustness of our approach across a larger spectrum of interaction modes, reinforcing the critical role of strategic fragmentation and MSA selection in achieving high prediction accuracy.
Our study provides a novel framework for the improved prediction of protein interaction sites involving IDRs. By harnessing the power of AlphaFold2-Multimer, adopting a strategic approach to protein fragmentation, and testing different techniques to concatenate the MSA of both proteins, we improve the modeling of protein interactions, opening up new possibilities for screening techniques to explore the vast landscape of protein networks within the cell.

T23

Folding-upon-binding pathways of an intrinsically disordered protein from a deep Markov state model

Thomas Sisk, Paul Robustelli[1]

[1]Dartmouth College, USA

A central challenge in the study of intrinsically disordered proteins is the characterization of the mechanisms by which they bind their physiological interaction partners. Here, we utilize a deep learning–based Markov state modeling approach to characterize the folding-upon-binding pathways observed in a long timescale molecular dynamics simulation of a disordered region of the measles virus nucleoprotein NTAIL reversibly binding the X domain of the measles virus phosphoprotein complex. We find that folding-upon-binding predominantly occurs via two distinct encounter complexes that are differentiated by the binding orientation, helical content, and conformational heterogeneity of NTAIL. We observe that folding-upon-binding predominantly proceeds through a multi-step induced fit mechanism with several intermediates and do not find evidence for the existence of canonical conformational selection pathways. We observe four kinetically separated native-like bound states that interconvert on timescales of eighty to five hundred nanoseconds. These bound states share a core set of native intermolecular contacts and stable NTAIL helices and are differentiated by a sequential formation of native and non-native contacts and additional helical turns. Our analyses provide an atomic resolution structural description of intermediate states in a folding-upon-binding pathway and elucidate the nature of the kinetic barriers between metastable states in a dynamic and heterogenous, or “fuzzy”, protein complex.

T24

Exploring the biophysical boundaries of protein families with deep learning methods

Miriam Poley-Gil[1], Alin Banka[2], Michael Heinzinger[2], Noelia Ferruz[3], Alfonso Valencia[1], R.Gonzalo Parra[1]

[1]Computational Biology Team, Life Sciences Department, Barcelona Supercomputing Center, Spain
[2]Department of Informatics, Bioinformatics & Computational Biology, Technical University of Munich, Germany
[3]Artificial Intelligence for Protein Design Group, Institute of Molecular Biology of Barcelona, Spain

The intricate interplay between protein sequence, structure and function remains one of the central questions in Molecular Biology. But recently, Machine and Deep Learning Models have revolutionised the field allowing us to explore the protein space faster. To understand what they are capturing and generating we have combined the use of state-of-the-art protein models for inverse folding (such as ProstT5 and ProteinMPNN) and for sequence generation (such as ProtGPT2 and ZymCTRL) with biophysical analyses focusing on the concept of local energetic frustration, to compare frustration signatures in a try to shed light on the evolutionary processes leading to the diversification of proteins. We have studied Globins, β-lactamases and also the subfamily RAS as preliminary examples.
The concept of local energetic frustration comes from the fact that proteins are optimised for folding and stability but also evolutionarily selected to function. We have developed a tool called FrustraEvo that measures the conservation of these local energetic frustrations within and between protein families (fully available on the server https://frustraevo.qb.fcen.uba.ar/). Moreover, we have made use of ESMFold to predict the structure of generated sequences and ProstT5 sequence embeddings to select the sequences.
We found that most of the highly frustrated native residues are related to functional aspects. These functional residues are mostly recovered by sequence generation models, which recover almost completely the native energetic signature suggesting that maybe there are alternative ways to design proteins instead of the way explored by evolution. In the case of catalytic sites, they are also recovered by inverse folding models. We therefore point out a selective memory concerning functionality highly influenced by the original training of the models, where they could have learnt co-evolutionary statistics. This may involve a primary level of memory (local). However, ProteinMPNN, despite this selective memory, seems an idealistic model minimising most of the highly frustrated positions but surprisingly generated diverse yet stable proteins and retained essential catalytic sites. Moreover, it also recovers the main network of frustrated contacts of the functional domains even suggesting a tertiary level of memory (contacts).
While evolutionary aspects remain yet unexplored extensively, our approach promises to effectively unravel the intricacies of protein family boundaries and explore design options for understanding protein evolution, including their implications in associated diseases. Ongoing works intend to implement this approach across all enzymes that have documented catalytic sites as well as relevant non-globular proteins.

T25

Perturbation modelling across single-cell omic modalities

Georgios Gavriilidis

In this presentation, recent advancements in single-cell perturbation modeling will be discussed, highlighting their transformative impact on cellular and molecular research. The effects of perturbations, i.e. external influences on cellular physiology, using technologies like CRISPR and compounds across various omic modalities, will be addressed. Additionally, a range of computational methods will be discussed, from classical statistical inference to machine and deep learning architectures, with an emphasis on the emergence of large foundational models inspired by language models. Finally, the challenges in single-cell perturbation modeling will be highlighted, and future directions proposed, including perturbation atlases, multi-omics datasets, and advancements in interpretability and performance using causal machine learning and multi-task learning.

T26

MAVISp, a modular structure-based framework for protein variant interpretation: current and future developments for supporting non-globular proteins

Matteo Tiberti[1], Matteo Arnaudi[1,2], Ludovica Beltrame[2], Kristine Degn[2], Mattia Utichi[1,2], Simone Scrima[1], Pablo Sánchez-Izquierdo Besora[1,2], Karolina Krzesińska[1,2], Alberte Heering Estad[1,2], Francesca Maselli[1,3], Terézia Dorčaková[1], Jordan Safer[4], Katrine Meldgård[1,2], Philipp Becker[2], Valentina Sora[1,2], Alberto Pettenella[1], Julie Bruun Brockhoff[1,2], Amalie Drud Nielsen[1,2], Jérémy Vinhas[1], Peter Wad Sackett[2], Claudia Cava[5], Anna Rohlin[6,7], Mef Nilbert[8,9], Sumaiya Iqbal[4], Matteo Lambrughi[1], Elena Papaleo[1,2]*

[1]Cancer Structural Biology, Danish Cancer Institute, 2100, Copenhagen, Denmark
[2]Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark, 2800, Lyngby, Denmark
[3]Institute of Molecular Bioimaging and Physiology, National Research Council (IBFM-CNR), Via F. Cervi 93, Segrate-Milan, 20054, Milan, Italy
[4]Bioinformatic and Computational Biology group, the Center for Development of Therapeutics, Broad Institute of MIT and Harvard, 415 Main St, Cambridge, MA 02142, USA
[5]Department of Science, Technology and Society, Scuola Universitaria IUSS, Istituto Universitario di Studio Superiori, Piazza della Vittoria 15, 27100 Pavia, Italy
[6]Department of Clinical Genetics and Genomics, Sahlgrenska university hospital, Sweden
[7]Department of Laboratory Medicine, Institute for Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
[8]Institute of Clinical Medicine, Dept. Oncology and Pathology, Lund University, Sweden
[9]Department of Clinical Research, Copenhagen University Hospital Amager and Hvidovre, Copenhagen, Denmark

The accumulation of genomic alterations over time, especially in tumor suppressor genes or oncogenes, is the prime cause of cancer. Such alterations influence disease development and outcome, risk of relapse and response to treatment to different degrees. While the number of known cancer-associated mutations is growing at an astonishing pace, the effect of most of them is currently unknown or uncertain (Variants of Unknown Significance, VUS). Attempting to bridge the gap, many computational methods that attempt to predict the effect of mutations have been devised; at the same time, structural biology and biomolecular simulations have also leapt forward thanks to improvements in hardware, methodology, and the availability of more accurate structure predictions. We introduce MAVISp [1] (Multi-layered Assessment of VarIants by Structure for proteins), a modular framework for the classification of the impact of variants affecting protein-coding genes and their corresponding protein products, based on the analysis of protein structures or protein structural ensembles. Our approach is based on high-throughput calculations of changes in folding free energy, biomolecular simulations of different types and other methods. In this contribution, we will highlight the current status and future work on MAVISp to support folded low-complexity regions and intrinsically disordered proteins or regions. For example, our framework currently includes predictions of short linear motifs and of the effect of mutations on their ability to bind target proteins. We will work towards adding support for dynamical properties of disordered regions that could be impacted by mutations. MAVISp currently includes data for more than 200 different proteins and tens of thousands of mutations, and is available at https://services.healthtech.dtu.dk/services/MAVISp-1.0/. We plan on expanding our dataset on a regular basis. We welcome displays of interest for new protein targets or contributors interested in using our approach in connection with their research.

[1] Matteo Arnaudi, Ludovica Beltrame, Kristine Degn, Mattia Utichi, Simone Scrima, Pablo Sánchez-Izquierdo Besora, Karolina Krzesińska, Alberte Heering Estad, Francesca Maselli, Terézia Dorčaková, Jordan Safer, Katrine Meldgård, Philipp Becker, Valentina Sora, Alberto Pettenella, Julie Bruun Brockhoff, Amalie Drud Nielsen, Jérémy Vinhas, Peter Wad Sackett, Claudia Cava, Anna Rohlin, Mef Nilbert, Sumaiya Iqbal, Matteo Lambrughi, Matteo Tiberti, Elena Papaleo. MAVISp: A Modular Structure-Based Framework for Genomic Variant Interpretation. bioRxiv 2022.10.22.513328; doi: https://doi.org/10.1101/2022.10.22.513328

T27

Classification and prediction of amyloid-forming propensity of amino acid sequences

Alisa Davtyan[1], Rafayel Petrosyan[1,2]

[1]L. A. Orbeli Institute of Physiology NAS RA, 0028 Yerevan, Armenia
[2]Zaven & Sonia Akian College of Science and Engineering, American University of Armenia, 0019 Yerevan, Armenia

Amyloids are stable aggregates of proteins rich in β sheets. They can stay for years in the human body and are associated with neurodegenerative diseases such as Alzheimer's disease, Parkinson's disease, and Huntington's disease, as well as with vital functions such as memory and the immune system. The accurate prediction of whether a given amino acid sequence will aggregate into an

amyloid is still an open question. The amyloid propensity and solubility prediction have been approached using statistical methods previously. We approach this problem with modern ML algorithms using increasingly available protein data. Our goal is to provide more efficient tools for classifying and predicting amyloid-forming proteins.

T28

Voronoi tessellation-based analysis of 3D conformations of non-globular proteins

Kliment Olechnovic, Sergei Grudinin[1]

[1]CNRS Laboratoire Jean Kuntzmann, Grenoble

There are multiple ways to computationally define and analyze interatomic contacts in a three-dimensional model of a molecule, but most studies in structural bioinformatics only rely on calculating and interpreting atom-atom distances. This is not surprising, as such analysis is usually fast and easy to implement. However, there are alternative approaches that are more descriptive. For example, the Voronoi tessellation and its variants were known for a long time and used to investigate various properties of molecular structures. Yet, such geometric methods have typically demanded substantial computational resources, hindering their practical application, especially for processing large structural ensembles that can come from modern experimental studies or from simulations. We will present a new, highly efficient method for computing tessellation-based atom-atom contacts constrained inside the solvent-accessible surface of a molecular structure. The method allows rapid analysis of multiple molecular conformations. Importantly, the efficiency of the new method is coupled with the practical ability of tessellation-derived contacts to describe interactions in a comprehensive manner, fully considering the relevant structural context and providing meaningful and summarizable interaction value descriptors - contact areas. We will also show how tessellation-derived contacts are utilized to define graphs that efficiently describe intermolecular interfaces, and how such graphs can be used to predict interface properties via attention-based graph neural networks.

SELECTED FLASH POSTERS

F01

Aggregation of LEA proteins: role in desiccation tolerance

Ana Pantelić[1], Dejana Milic[1], Milan Senćanski[1], Zuzana Bednarikova[2], Marija Vidović[1]

[1]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Serbia
[2]Department of Biophysics, Institute of Experimental Physics, Slovak Academy of Sciences, Kosice, Slovakia

Drought is one of the greatest threats to global food security and poses a major challenge for agriculture. As a result of climate change, the frequency and intensity of droughts are increasing, necessitating the development of drought-resistant crops. Resurrection plants, able to survive prolonged periods of desiccation and to fully resume metabolic activities shortly after re-watering, present excellent models for studying desiccation with the potential to improve the crop drought tolerance of crops.
A hallmark of desiccation tolerance in resurrection plants is the accumulation of intrinsically disordered, late embryogenesis abundant (LEA) proteins. Their high structural plasticity allows them to interact with various ligands and partners and stabilise membranes and enzymes during water deficit. In addition, they can form intracellular proteinaceous condensates that increase structural integrity and intracellular viscosity during drying. However, no specific cellular targets of LEAPs have been identified so far.
We recombinantly produced desiccation-induced LEA4 protein from the resurrection plant Ramonda serbica and investigated its oligomerisation potential. Size exclusion chromatography pointed out that selected LEA4 protein is prone to aggregation. FTIR studies showed that the selected LEAP sample is composed of 34% α-helical or random structure, 26 % of β-turns and 26 % of β-sheets, suggesting a low probability for amyloid protein organisation. In addition, the aggregation pathway was studied with amyloid-specific dye, thioflavin T, and with dye sensitive to hydrophobic surfaces of proteins (ANS). Atomic force microscopy showed that the produced LEA4 protein sample contains many aggregates of various sizes, from small (15-25 nm) to bigger (approx. 35 nm in size) and a few large aggregates with heights around 100 nm. Finally, the obtained results indicated a non-amyloid aggregation pattern of LEA4 protein.
Our results are important for further elucidation of protective LEAP's mechanism during desiccation at the cellular level. Structural flexibility and aggregation propensity of LEAP might be crucial in direct, or indirect (e.g. via LLPS) maintaining of the native conformations of biomolecules (proteins, membranes and DNA). We suggest that native or bioengineered LEAPs might be considered as target molecules for the development of strategies to improve the drought resistance of crops.
Keywords: aggregation, intrinsically disordered proteins, late embryogenesis abundant (LEA) proteins, Ramonda serbica, resurrection plants

Acknowledgements: This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia (Contract No. 451-03-47/2023-01/ 200042) and by the Science Fund of the Republic of Serbia-RS (PROMIS project LEAPSyn-SCI, grant no. 6039663).

F02

Homorepeat variability within the human population

Pablo Mier[1], Miguel A. Andrade-Navarro[1], Enrique Morett[2]

[1]JGU Mainz, Germany
[2]UNAM, Mexico

Population genetic variation is a driving force in evolution. Observed genetic variation is in general neutral since deleterious alleles are readily discarded, although slightly deleterious ones can survive by a range of different mechanisms, such as linkage to strongly selected alleles. Human genetic variation can now be studied in depth thanks to the large amount of data collected in multiple genomic projects, allowing us to determine which genetic changes are silent or with small phenotypic consequences and which are strongly conserved and therefore much more likely to be involved in critical functions. Unlike the average protein sequence, homorepeat evolution may be driven by DNA replication slippage, caused by nucleotide repeats where DNA polymerases can add or skip repeats. DNA replication slippage therefore changes the length of homorepeats. There are several diseases that result from pathogenic changes in homorepeat length, however, naturally occurring length variation has been little explored. In this work, we studied the population variation of homorepeat length and found extraordinary variation, implying homorepeats as mutational hotspots. Glutamine homorepeats (polyQ) were studied in depth and showed length-dependent variation: shorter polyQ tend to vary widely, while large polyQ experience primarily deletions. Interestingly, more evolutionarily conserved polyQ tend to vary less, implying stronger functional constraints. Our results show that there is a large variation in the length of homorepeats in the human population, probably without phenotypic consequences.

F03

Jumpcount: Exact confidence intervals for free energy difference estimations

Jan Beránek[1], Pavel Kříž[2], Vojtěch Spiwok[1]

[1]Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague, Czech Republic
[2]Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic

Increasing number of biomolecular simulations provide sufficiently long simulations to predict equilibrium constants and associated free energies of simulated processes such as protein folding or protein-ligand binding. The equilibrium constant of a process of a transition from the state A to state B can be estimated as the fraction of times spent in B and A. Here we present a simple, yet exact, method to calculate the errors of the free energy estimations, which is calculated solely from temperature and the number of transitions from A to B and B to A. Markovianity of the process is the prerequisite of the method. The error can be calculated online by our tool at jumpcount.cz.
This method can be used for results based on unbiased molecular dynamics simulations or simulations based on enhanced sampling methods with static bias potential, e. g. umbrella

sampling. The approach was tested on numerous molecular systems (glycerol, alanine dipeptide and miniprotein folding).
The work was supported by Czech Science Foundation (22-29667S and 19-16857S) and the Ministry of Education, Youth and Sports of the Czech Republic (LM2018140, LM2018131).

F04

Predicting disordered proteins in Zea mays L.

Dragana Dudić[1], Bojana Banović-Đeri[2], Ana Nikolić[3], Jelena Samardžić[2], Jelena Vančetović[3], Nenad Delić[3]

[1]Faculty of Informatics and Computer Science, The University "Union-Nikola Tesla", Belgrade, Serbia
[2]Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[3]Maize Research Institute, "Zemun Polje", Belgrade, Serbia

Introduction.
Intrinsic protein disorder is a phenomenon characterized by a lack of defined three-dimensional structure of a fully functional protein with characteristically low complexity and bias in amino acid composition. Recent studies have shown that several intrinsically disordered proteins are associated with plant response to abiotic stress which has serious effects on plant development.
A Zea mays L. is a worldwide important crop with its yield negatively affected by a wide range of abiotic stresses and adequate gene annotation is crucial for understanding the plant response. Due to the repetitive nature of the maize genome, the process of gene annotation lacks accuracy and accurate protein structure prediction can improve this process.
Materials and Methods.
Our study involves maize leaf total transcriptome data obtained via Illumina MiSeq platform and 150bp paired-end sequencing for 46 inbred lines of Zea mays L. from Maize Research Institute "Zemun Polje" that can be classified as Lancaster or non-Lancaster lines.
Transcripts are quantified using Kallisto software and gene level analysis is conducted using DESeq2, EdgeR and limma R packages. Differentially expressed genes were annotated using BLAST2GO and validated with qPCR.
Protein structure predictions were made and compared using AlphaFold2 and ESMFold tools, available at MaizeGDB website and Foldseek as a new method with MMseqs2 for multiple sequence alignment as an optimization step.
Results.
We identified 77 differentially expressed genes between Lancaster and non-Lancaster groups grown under optimal conditions. Cold and drought tests are conducted on the two subsets, both consisting of four Lancaster and four non-Lancaster lines.
Obtained results include transcripts for 27 protein coding genes as significant for maize drought and cold tolerance all of which are associated with intrinsically disordered proteins. For selected genes we identified all their isoforms and predicted protein structure. Results showed that alternative isoforms, functionally based on structure, are more suitable than reference ones available at MaizeGDB database.
Conclusion.

Intrinsically disordered proteins play an important role in Zea mays L. adaptation to abiotic stress through improvement of annotation of the protein coding genes using their structural prediction.

F05

Collective Classification Algorithms in Identifying Intrinsically Disordered Proteins within Protein-Protein Interaction Networks

Milana Grbić, Nenad Vilendečić, Milan Predojević and Dragan Matić[1]

[1]Faculty of Natural Sciences and Mathematics, University of Banja Luka

In the realm of molecular biology, deciphering the intricacies of protein structures is vital for unraveling cellular processes and disease mechanisms. Intrinsically disordered proteins (IDPs) play crucial roles in various cellular functions, from signaling cascades to molecular recognition events. However, identifying IDPs amidst a vast protein landscape presents a formidable challenge due to their dynamic nature and elusive characteristics. This study aims to address this challenge by employing Collective Classification Algorithms (CCAs) on Protein-Protein Interaction (PPI) networks to identify IDPs.
Our research focuses on evaluating the effectiveness of several CCAs in identifying IDPs within the human PPI network. By examining relationships between a protein and its neighboring nodes, we seek to uncover patterns indicative of intrinsic disorder.
Our application of CCAs anticipates two main outcomes. Firstly, we expect these algorithms to accurately pinpoint IDPs within the entire PPI network. Secondly, by applying CCAs to pre-clustered network components obtained through community detection algorithms, we aim to achieve precise and accurate identification of IDPs within specific functional modules. Successful outcomes could advance our understanding of IDPs' functional relevance and facilitate targeted investigations into their roles in specific cellular pathways and disease contexts. Through this research, we aim to provide valuable insights that may drive progress in molecular biology and inform the development of therapeutic strategies targeting intrinsically disordered proteins.

F06

Side-chain propelled rings (SPURs) in globular and disordered proteins

Rostislav Skrabana[1], Katarina Martonova[1], Radovan Dvorsky[1], Michal Hricovini[2], Milos Hricovini[2], Ondrej Cehlar[1]

[1]IN SAS, Slovakia
[2]ICH SAS, Slovakia

Local intramolecular rings are the smallest secondary structure elements found in PDB. These rings are held together by one or more hydrogen bonds, their length may range from eight up to 20 atoms. Rings involving exclusively main chain atoms lead to bends in the protein backbone, forming

α-, β-, gamma-, delta- or pi-turns. Even more variable are rings exhibiting side chain/main chain interaction. We have coined a nick name SPURs (Side-chain Propelled Rings) for the latter.
The structure of globular proteins contains large number of regularly repeating intramolecular hydrogen bonds between main chain atoms, which create secondary structure elements, e.g. β-sheets or α-helices. SPURs and main chain turns are frequently encountered at the borders of these larger secondary structure types. The situation in non-globular, intrinsically disordered proteins (IDPs) is different. IDPs contain no stable three-dimensional structure, therefore IDPs' polypeptide chain is hydrated and its conformations are composing a conformational ensemble (CE). Individual conformational states of CE are separated by a small energy barrier and freely interconvert [1]. IDPs satisfy hydrogen-bonding potential of their main chain atoms almost exclusively by hydrogen bonding to the solvent molecules.
We wondered if small hydrogen-bonded rings would occur in IDPs as well, despite prevalent propensity to H-bonding with water. Using experimental and computation approaches we found an indication of intramolecular SPURs in IDPs. Results of simulations show that the lifetime of these rings may be short but they are frequently renewed. Quantum mechanics calculations revealed the proportion of rings in the local CE.
It is hypothesized that the conformational restrains inherent to small rings influence the composition of IDP CE and regulate corresponding biological activity. We have observed several SPURS in crystals of IDP tau with monoclonal antibodies [2]. In the future work we aim to examine the SPURS also in other IDPs.

1. Thomasen FE, Lindorff-Larsen K. Conformational ensembles of intrinsically disordered proteins and flexible multidomain proteins. Biochem Soc Trans. 2022 Feb 28;50(1):541-554. doi: 10.1042/BST20210499. PMID: 35129612.
2. Cehlar O, Bagarova O, Hornakova L, Skrabana R. The structure of the unstructured: mosaic of tau protein linear motifs obtained by high-resolution techniques and molecular simulation. Gen Physiol Biophys. 2021 Nov;40(6):479-493. doi: 10.4149/gpb_2021031. PMID: 34897022.

This work was supported by APVV-21-0479, VEGA 2/0141/23, 2/0125/23 and EU Grants ADDIT-CE, InterTau.

F07

Dynamic and Energetic Consequences of Disulfide Bonds in Proteins

Miguel Fernandez-Martin, R. Gonzalo Parra[1]

[1]Barcelona Supercomputing Center, Barcelona, Spain

Disulfide bonds or bridges are a type of covalent bonds formed through the oxidation of cysteine residues, traditionally known for their role in stabilizing protein structures. However, recent research starts to show that disulfide bonds can be more versatile than thought. Here are 3 different scenarios: 1) MCoTI-II, a cyclotide trypsin inhibitor, where disulfide bridges contribute to the formation of a cyclic cystine knot (CCK) structural motif. Molecular dynamics simulations highlight a delicate interplay of frustration and correlation among specific disulfide bridges during the folding process.

Notably, the Cys4-Cys21 bridge exhibits functional flexibility, suggesting its potential significance in the peptide's biological activity. 2) in the N-terminal domain of the oxidoreductase DsbD (nDSBd) the disulfide bonds introduce frustration, influencing the conformational dynamics of specific residues. This frustration mechanism is conserved across bacterial species, showcasing the regulatory role of disulfide bonds in functional signaling. 3) in Azurin, although it has been shown that thermal and chemical stability are dramatically reduced when the disulfide bond is absent, the protein adopts a folded structure identical to that with an intact disulfide. The Cys3-Cys26 was shown to modulate the native state stability but also the transition state ensemble.
In this work, we aim to study the different types of disulfide bonds that are present in the Protein Data Bank (PDB). We have analyzed the local energetic frustration patterns around these covalent bonds to understand how disulfide bonds can either locally stabilize the structure of a protein or, in contrast, introduce local instabilities linked to functional reasons. Our analysis includes all the structures present in the PDB ranging in length from 20 to 600 residues with at least one disulfide bond (full dataset n=11158), as well as a subset of proteins with experimental structures in both oxidized and reduced states for at least one disulfide bond (paired dataset n=1053). Using the Frustratometer algorithm, we assessed local frustration patterns across both datasets and complemented it with the analysis of predicted intrinsic disorder using IUPRED.
Our work sheds light into understanding the different roles that disulfide bonds play in biological systems. We have classified our dataset into different scenarios to provide an interpretable landscape of how disulfide bonds alter the dynamic and energetic properties of proteins. Further understanding disulfide bonds is of relevance as being able to induce their reduction/oxidation could be used to trigger specific cellular responses and be used as biomedical modulable devices.

F08

Structural ensembles of the single alpha-helix (SAH) segment of the postsynaptic protein Drebrin

Soma Varga[1], Bálint Ferenc Péterfia[1], Perttu Permi[2], Zoltán Gáspári[1]

[1]Faculty of Information Technology and Bionics, Pázmány Péter Catholic University, Budapest, Hungary
[2]Department of Biological and Environmental Science, Nanoscience Center, University of Jyväskylä, Finland

Drebrin (developmentally regulated brain protein) is scaffold protein participating in the organization of the postsynaptic density (PSD). It establishes physical contact between other PSD scaffold proteins like the Homer family and F-actin. It has been proposed to play a key role in the development and reorganization of dendritic spines, contributing to neural plasticity. Besides an N-terminal globular ADF-H (Actin-Depolymerizing Factor Homology) domain, Drebrin contains large regions that are predicted to be intrinsically disordered. Between these, the segment 173-238 is predicted to form a single-alpha helical (SAH) domain. To analyze the structural and dynamical properties of this region, we have expressed and purified the SAH segment and conducted investigations, with CD, NMR and SAXS. Resonance assignment of the region required the use of 4D experiments to overcome severe signal overlap resulting from the low-complexity and repetitive

nature of the underlying sequence. Our experiments confirm that the region exhibits a stable helical structure and is monomeric in solution. Using computational tools developed in our group, DIPEND and CoNSEnsX+, we have generated structural ensembles consistent with secondary Cα and Cβ chemical shifts, as well as SAXS data. The ensembles reflect a largely rodlike structure with moderate deviations from a straight, continuous helix, with differently decreasing helicity at both termini.

F09

Systematic analysis of protein interactions using PrISMa

Marta Mendes, Daniel Perez-Hernandez, Sophie Rodius, Gunnar Dittmar[1]

[1]Luxembourg Institute of Health

Protein-protein interactions are an important part of the post-translational regulatory layers which regulate cellular growth. Besides protein interactions, which are mediated by fully folded domains, intrinsically disordered regions (IDRs), regions without a defined fold, have gained recognition as important sites for protein-protein interactions and as regulatory stretches in a protein. IDRs are also the target of many post-translational modifications, which regulate the interactions mediated through these regions. Although significant efforts have been undertaken to describe protein-protein interaction networks, the interaction site has been mapped only for a small group of protein-protein interactions.
The PrISMa method allows to solve this problem as it can determine all SLiM-mediated interactions in a single mass spectrometry-based experiment. By including post-translational modification in the screening matrix, we were able to identify several new PTM-regulated interactions. These included regulations by phosphorylations but also the rarer modifications of arginine methylation or citrulination. By systematically analysing the C-terminal tails of all members of the human claudin family, we identified new common interactions with the proteasome and the CCT complex for a subset of the claudin family. A similar screen of the C/EBP transcription factor family identified several common interaction sites, which indicated a significant cross-talk between the different members of this protein family.

F10

Exploring Disease-Associated Implications of Non-Synonymous Single Nucleotide Polymorphisms in the Human Fibronectin 1 Gene: A Comprehensive Analysis of Functional and Structural Consequences

Mirela Lika (Çekani)[1], Xhiliola Bixheku[2], Anila Hoda[3]

[1]University of Tirana, Albania
[2]Quality Assurance Agency for Higher Education, Albania
[3]Agricultural University of Tirana, Albania

This study represents a comprehensive investigation into the functional and structural implications of non-synonymous single nucleotide polymorphisms (nsSNPs) in the human Fibronectin 1 (FN1) gene (ENSG00000115414). FN1 plays a critical role in cellular adhesion, migration, and signaling, making understanding the consequences of genetic variations crucial for elucidating potential associations with various disorders. Leveraging state-of-the-art bioinformatics tools, we identified thirteen nsSNPs within the FN1 gene, characterizing them as deleterious based on analyses using SIFT, PolyPhen2, SNAP2, and Panther. Notably, these mutations exhibited significant variations in tolerance indices, with certain variants being classified as highly deleterious, emphasizing their potential impact on protein function.

To assess the stability implications of these nsSNPs, we employed predictive algorithms such as DDG I-Mutant and DDG Mupro, consistently revealing a destabilizing effect on FN1 protein stability. Subsequent structural analyses provided a nuanced understanding of the mutations' effects on FN1 dynamics, compactness, solvent accessibility, and flexibility. MutPred2 predictions further underscored structural alterations, emphasizing the potential consequences for FN1 functionality. Evolutionary conservation analysis and phylogenetic studies added a layer of insight, shedding light on the conservation patterns and evolutionary relationships of the FN1 gene across species.

Molecular dynamics simulations were employed to elucidate changes in FN1 stability induced by the identified nsSNPs, while Prodigy analyses provided valuable data on the binding affinities in the context of protein-protein interactions. STRING analysis unraveled the intricate FN1 protein-protein interaction network, emphasizing the gene's integral role in fundamental biological processes.

Exploring the impact of these mutations on FN1 interactions with other cellular components, revealed alterations in binding energetics. Post-translational modification predictions and additional molecular dynamics simulations offered a comprehensive view of the functional consequences of these nsSNPs.

This comprehensive approach significantly advances our comprehension of the molecular and phenotypic repercussions stemming from deleterious non-synonymous single nucleotide polymorphisms (nsSNPs) within the human FN1 gene. It yields pivotal insights crucial for discerning diseases associated with FN1, presenting a promising avenue for unraveling FN1-related disorders. Furthermore, the findings pave the way for tailoring therapeutic interventions based on individual genetic profiles.

Keywords: Fibronectin 1, nsSNPs, genetic variations, protein stability, structural consequences, molecular dynamics simulations, protein-protein interaction network, evolutionary conservation.

F11

Evaluating AlphaMissense Mutation Pathogenicity in Intrinsically Disordered Proteins using DisCanVis

Norbert Deutsch, Gábor Erdős, Mátyás Pajkos, Zsuzsanna Dosztányi

Genome projects have led to the accumulation of a vast number of genetic variants, crucial for understanding the genetic basis of diseases like cancer. Despite this, most of these variants' clinical significance remains unknown, posing a challenge in classifying them, particularly in multigenic

diseases. Pathogenic mutations often occur in critical regions of proteins, affecting their function and structure, and thus contributing to disease pathogenicity. A recently published machine learning-based method, AlphaMissense, predicts the pathogenicity of mutations. Although mutations in disordered regions were also examined, the method was primarily focused on categorizing mutations in structured regions and its effectiveness in disordered regions remains uncertain.

Our primary aim was to characterize the effectiveness of the AlphaMissense method when applied to disordered regions, in particular, if it is suitable for predicting pathogenicity in disordered regions, and whether it can identify functional areas within these regions. In order to analyze the impact on disordered regions on a large scale, we used DisCanVis, our web server that links structural and functional annotations with genomic data. We used this platform as a foundation to which we integrated pathogenicity scores predicted by AlphaMissense. Using this system, we conducted detailed analyses of the pathogenicity values of functional units within disordered regions, focusing particularly on Short Linear Motifs (SLiMs) and their sequential environments.

Our analysis using AlphaMissense suggests that mutations in disordered regions are often categorized as non-pathogenic. It appears that the pathogenicity of Short Linear Motifs (SLiMs) may be related to their conservation pattern, resembling 'Island Conservation,' where motifs are more conserved than their adjacent sequences. This observation could indicate a 'Pathogenicity Hill' effect, where SLiMs might display higher pathogenicity compared to their less conserved neighboring sequences. Although AlphaMissense is primarily focused on structured proteins, our findings imply that it could also identify conserved areas within disordered regions, potentially underlining their functional significance.

These results suggest that in conjunction with DisCanVis, AlphaMissense could serve as a valuable supplementary tool for analyzing mutations in proteins with less well-defined annotations.

POSTERS

P01

Molecular dynamics study on evolutionary loss of calcium binding in beta-gamma crystalline proteins

Agostina Spiga, Galo Ezequiel Balatti, Sebastián Fernández Alberti, María Silvina Fornasari[1,2]

[1]Department of Science and Technology, National University of Quilmes, Buenos Aires, Argentina.
[2]National Council for Scientific and Technical Research (CONICET), Argentina.

Introduction.
Beta gamma crystallins are the predominant structural proteins in vertebrate eyes. They are involved in the development of ocular abilities and also have several mutations already described that lead to ocular pathologies, such as the formation of cataracts. In vertebrates they are monomeric proteins with two globular domains, while in microorganisms or some invertebrate eukaryotes they are monomeric with a single domain. In the transition from single domain to double domain, these proteins lost the ability to bind calcium, for some explanation that we believe is entropic based. A biophysical study together with an evolutionary study would serve to understand how the increase in stability and low affinity for calcium occurred in vertebrates. In this work, we started from structural alignments and chose three proteins with apo and holo forms and three more that do not bind calcium. Of the first three, two belong to bacteria and one to an invertebrate eukaryote. Of the last three, one belonging to the fungus, another to the bacteria and another to a mammalian vertebrate . We visualize that residues that interact with calcium and their equivalents in proteins that do not bind it fall in flexible regions, and that calcium-binding motifs are not recognizable in the lens of the human eye. They were carried out molecular dynamics, clustering analysis and study of RMSD and RMSF for each case.
Results.
Firstly, we found that in proteins that present apo and holo forms, a greater number of clusters is visualized in the apo forms.
Secondly, the apo forms of these proteins have higher RMSF than the holo forms.
Third, we found that apo forms of calcium-binding proteins have higher overall RMSD than non-calcium-binding proteins.
Conclusion.
These results demonstrate for the first time molecular dynamics studies carried out on key species on an evolutionary scale, which are in line with experimental studies and which demonstrate that beta crystallines in vertebrates lost their affinity for calcium, compensating for structural rigidity at the expense of their primary structure.

P02

A comprehensive exploration of human tandem repeat RNA binding proteins: synergies between sequence and structure analysis

Agustín Ormazábal, Matías Sebastián Carletti, Tadeo Enrique Saldaño, Martín Gonzalez Buitron, Julia Marchetti, Nicolas Palopoli, Alex Bateman

The unique arrangements of protein repeats provide a scalable solution to bind, read and process RNA sequences effectively, where each repeat could bind to one or a few ribonucleotides. Moreover, many tandem repeats in proteins were recently associated with Pfam families as those involved in transcription, RNA-splicing or stabilisation, and multiprotein complexes assembly. However, studies characterising tandem repeat RNA-binding proteins are lacking compared to widely explored interactions between repeat proteins and DNA. Due to the availability of high-quality structure predictions based on IA for nearly all human proteins, a comprehensive census of RNA-binding repeat proteins is now timely. We performed a large-scale analysis of human proteins devoted to expanding the knowledge about tandem repeat proteins experimentally reported as RNA-binding molecules. A total of 219 tandem repeat proteins that bind RNA molecules were detected and characterised. The overlap between repeat regions and RNA-binding motifs was analysed as a first step towards assessing their functional relationship. We observed differences in the characteristics of repeat regions predicted by sequence-based or structure-based methods according to their degree of disorder, sequence composition, functions, and protein domains. The results showed that sequence and structural methods' combination finds significantly more tandem repeat proteins than either method alone. Finally, we present examples of novel tandem repeat RNA-binding proteins.

P03

Identifying function of Intrinsically disordered proteins (IDPs) with hyperbolic mapping of human protein-protein interactome

Aimilia Chrstina Vagiona, Miguel Andrade Navarro[1]

[1]Johannes Gutenberg University Mainz, Germany

Protein–protein interactions (PPIs) form a complex network called “interactome” that regulate many functions in the cell. In recent years, there is an increasing accumulation of evidence supporting the existence of a hyperbolic geometry underlying the network representation of complex systems such as the interactome. In particular, it has been shown that the embedding of the human Protein-Interaction Network (hPIN) in hyperbolic space (H2) captures biologically relevant information. More specifically, the radial positioning of nodes encapsulates information about the conservation and the evolution of proteins, while their angular positioning captures the functional and spatial organization in the cell. Additionally, intrinsically disordered proteins (IDPs) and regions (IDRs) lack a well-defined and folded three-dimensional structure in the absence and/or presence of a binding partner. Here we embedded the hPIN in the two-dimensional hyperbolic plane using the R package “NetHypGeom”, which implements the LaBNE + HM algorithm. This algorithm combines manifold learning and maximum likelihood estimation to uncover the hidden geometry of complex networks. Here we explore whether the mapping of hPIN into the H2 and the clustering based on the angular dimension that contains functional organization in the cell, would allow us detect the

function of IDPs/IDRs. Also, we show that within protein interaction networks, IDPs often represent hubs carrying multispecific binding sites that adapt to different binding interactors. Functional enrichment analysis of the IDPs while the interactors of IDPs can lead us to a better understanding of the intrinsically disordered proteins.

P04

The role of distinct lipid compositions in the initial membrane-binding of tau protein

Catarina Pimenta[1], Joana Paixão[1], Ana Coutinho[1,2], Ana Azevedo[1], Manuel Prieto[1], Ana M. Melo[1]

[1]Institute for Bioengineering and Biosciences and Associate Laboratory i4HB, Instituto Superior Técnico, ULisboa, Portugal
[2]Dep. Química e Bioquímica, Faculdade de Ciências, ULisboa, Portugal

The pathological deposition of the intrinsically disordered protein tau in the brain is a hallmark of numerous tauopathies, including Alzheimer's disease. Accumulated evidence supports that lipid vesicles can induce tau misfolding and self-assembly, and subsequently tau oligomers promote membrane permeabilization. Moreover, recent studies revealed that hyperphosphorylated tau is unconventionally secreted by direct translocation across the plasma membrane. Specifically, this release comprises an initial membrane localization and clustering of tau, which was suggested to be highly dependent on cholesterol (Chol) and sphingomyelin (SM) ("raft" lipids), and also in phosphatidylinositol 4,5-bisphosphate [PI(4,5)P2] lipids. However, in vitro tau-lipid interaction studies have focused so far on MTBR-fragments (lacking both termini) or PHF6 hexapeptides. Here, we aim to obtain a detailed understanding of the initial steps of membrane-binding of 2N4R tau protein (longest isoform). We employed fluorescence approaches based mainly on fluorescence correlation spectroscopy (FCS) to quantify the contributions of electrostatic effects and raft-mimicking compositions on tau-lipid interaction. We explored anionic POPS or PI(4,5)P2-contaning lipid vesicles, and also a 1:1:1 POPC:Chol:SM ternary mixture with co-exiting liquid ordered /disordered membrane phases. Our FCS data support that tau binds preferentially to negatively-charged liposomes with either POPS or PI(4,5)P2 lipids, revealing that tau-lipid interaction is not dictated by a specific binding to PI(4,5)P2 but instead is mainly governed by electrostatic effects. Together, our work highlights the role of distinct lipid contents /membrane physical properties for tau recruitment to the membrane surface.
Work supported by FCT-Portugal (2022.01454.PTDC grant and CEECIND/00884/2017 contract to AMM, and UIDB/04565/2020 and LA/P/0140/2020 funding to iBB and i4HB) and Maria de Sousa Prize- 2nd edition funded by Bial Foundation and PT Medical Association (grant 45/2022 to AMM).

P05

The major components of cerebrospinal fluid alters inhibition and aggregation pathway of amyloid-β and α-synuclein

Andrius Sakalauskas, Arune Jocyte, Ruta Snieckute, Ziaunys Mantas, Smirnovas Vytautas[1]

[1]Life Sciences Center, Vilnius University, Vilnius, Lithuania

Protein aggregation into fibrillar aggregates is commonly related with different neurodegenerative disorders, including Alzheimer's (AD) and Parkinson's (PD) diseases. The most common neurological condition is recognized to be AD that affects over 35 million people worldwide with the number forecasted to double by 2030. The search for anti-amyloid-β drugs to slow-down or prevent the disease progression is a challenging task. The primary objective is to determine the active molecule, which is commonly achieved by in vitro screening experiments.
While there are multiple reports showing potential drugs, they often provide aggregation data conducted in PBS or phosphate buffer solutions with varying ionic strengths and pH conditions. However, such data does not resonate with the actual conditions where the Aβ, α-synuclein or other disease related proteins aggregate in human brain. Hence, it is crucial to test target molecules in media that closely resembles the conditions in the brain. One of the possible ways is to use major components of cerebrospinal fluid (CSF) that may dictate the protein aggregation pathway and the efficacy of anti-amyloid molecules.
In this research, we composed an artificial medium that mimics artificial cerebrospinal fluid (aCSF) and tested how Aβ and α-synuclein aggregates in such medium comparing to standard PBS conditions. Further, we questioned how components in aCSF affect inhibitory potential of known inhibitors such as oxidized EGCG and fluorinated benzenesulfonamide (VR16-09). Our findings demonstrate that major components of CSF alter both protein aggregation kinetics and morphology of formed fibrils. This alteration appears to result from to the combined effect of multiple components, such as HSA, glutamine, cholesterol and others. In addition to that, altered aggregation pathway affected the inhibitory potential of both tested molecules, leading to a significant change between the results and conclusions regarding their anti-amyloid potential.

P06

Panoramic view of ML algorithms and their use in Tandem Repeat Protein tools

Stefany Neciosup-Vera, Layla Hirsh[1]

[1]Department of Engineering, Pontificia Universidad Católica del Perú, Lima, Perú

Machine Learning is an IA area that uses data to develop algorithms for discovering trends and patterns. A novel example of an ML approach is AlphaFold, which incorporates physical and biological knowledge about protein structure, leveraging multi-sequence alignments, into the design of the deep learning algorithm [1]. As ML is becoming more popular, many questions arise: "Which algorithm is the right one by aim and features type?", "How to differentiate between categorical and numerical features?", "Which is the minimum amount of data and features needed?", "What are the summary and association measures by feature type?". The first step is to differentiate input(X) from output(Y), considering a function defined as f(X), affected by the value of X, let Y = f(X), and f(X) = b + mX, if X takes on the values {4, 5}, assuming b=3 and m=2, Y takes on the values {11, 13} respectively. Here, we see that Y (output) is affected by X (input), and both Y and X are numerical features, but they could be categorical or mixed features. For both being categorical features, let X={0,1,2,3,4,5} as Structural Repeat Class (SRC) with 0 as none of those, let Y={repeat, no repeat}, and

f(X)="Repeat" if X>=1, f(X)="No Repeat" if X=0. Here input should be defined as categorical, because it is not possible to operate between X values. For a mixed features case, let numericalX1=TM-score, numericalX2=Sequence identity between two protein structures (A and B), and categoricalX3=SRC of proteinA. Taking into account that repeat proteins have a degenerated sequence but they conserve a structural similarity, we could define a categorical output as Y=Belonging of proteinB to the same SRC of proteinA. Now we should decide which is the right model for each of our examples. With the aim of help in this decision, we have done an exhaustive review of literature, and created a detailed diagram based on the objective of the analysis (describe, predict or optimize), and the corresponding methods (clustering, dimension reduction, regression, classification, other), by the type of inputs and outputs as we describe before. Furthermore we list, if there exist, the related repeat protein tool(s). We hope that using this diagram, the selection of the ML method could be less blurry, by having a panoramic view with the most known possibilities.

P07

Predictive Modeling of Non-Globular Proteins Using Mixed Linear Models with Fixed and Random Factors

Dário Ferreira, Sandra Ferreira[1]

[1]University of Beira Interior, Portugal

The analysis and classifications of non-globular proteins are central elements in today's bioinformatics research. This presentation discusses an innovative approach using mixed linear models combining fixed and random effects factors, providing an ideal framework for modeling the complex structure of non-globular proteins.
The methodology includes estimating parameters and variance components, which are crucial to the model's predictive accuracy. A range of popular methods such as ANOVA, ML, REML, and Bayesian is utilized. At the same time, the approach is not constrained by the common assumption that the random part follows a normal distribution, therefore allowing more flexibility for the unusual distributions often encountered in the study of non-globular proteins.
We consider linear models with fixed vectors and random vectors, which can accommodate different continuous distributions. This model effectively addresses the challenging expectation of non-normal distributions often seen with non-globular proteins.
The approach provides estimators for variance components and future observation predictions. If our model's assumptions hold, we can produce unbiased estimators, offering an innovative link between the complex structure of non-globular proteins and linear modeling.
In terms of prediction, we apply the Least Square Estimators (LSE), generate a sample to replicate future observations' behaviour, and provide limit level confidence intervals.
By applying the inverse of the Glivenko-Cantelli theorem, we calculate the empirical quantiles that represent the confidence interval for future random vectors. This capacity offers a novel method for testing hypotheses on future observations.
Our comprehensive modeling framework provides a detailed understanding of the intricate nature of non-globular proteins. It is poised to serve as a basis for future integration of more complex algorithms for the advancement of bioinformatics in the study of non-globular proteins.

P08

Optimizing Disordered Flexible Linker (DFL) Analysis: A Comprehensive Domain-Linker-Domain Dataset Collection Approach

Di Meng[1], Juliana Glavina[2], Gianluca Pollastri[1], Lucía B. Chemes[2]

[1]School of Computer Science, University College Dublin, Belfield, Dublin 4, Ireland
[2]Instituto de Investigaciones Biotecnológicas (IIBiO-CONICET), Universidad Nacional de San Martín, Av. 25 de Mayo y Francia, CP1650 Buenos Aires, Argentina

Disordered flexible linkers (DFLs) are a subclass of intrinsically disordered regions (IDRs) that carry out functions closely linked to their conformational flexibility. DFLs play key roles in connecting domains and maintaining inter-domain interactions by linking two functional regions. Despite their prevalence in most proteins, only approximately 300 DFLs have been experimentally annotated. Our primary objective is to compile a large, high-quality, and comprehensive DFL dataset suitable for training machine learning algorithms, including Deep Neural Networks. This project aims to enhance our understanding of DLFs and develop highly accurate predictors for DFL functionality.
We propose a three-class classification scheme for DFLs, including DFLs connecting two independent domains (DLD, or IDL-independent domain linker), a motif and a domain (MLD/DLM), or two motifs (MLM). We collected more than 1900 DLD-IDL that connect independently folded domains (DLD-SCOP) and conducted an in-depth analysis of their distinctive features. Leveraging a protein language model (protTrans), we embed protein sequences to extract and scrutinize features specific to DLD-IDLs, distinguishing them from dependent-domain linkers, intradomain loops, and other disordered regions. Moreover, we have implemented an automated updating mechanism for this process, ensuring continual enhancement of the dataset. This approach holds promise for training machine learning models on a broader and more comprehensive dataset compared to existing resources like Disprot Linkers.
Bioinformatics algorithms offer a viable means to gather DFLs, enabling the annotation of novel instances and fostering a deeper understanding of these linkers. Subsequently, we aim to construct an initial predictor utilizing our current dataset and benchmark its performance against existing DFL predictors. We will then delineate data collection protocols to extract DLD-DFLs and MLM-DFLs from the Alphafold database and the ELM database, respectively. This will allow us to perform further analysis on an enriched and high-quality DFL database containing DLD-SCOP, DLD-Alphafold, MLM-ELM, and Disprot linkers.

P09

Structural Dynamics and Pharmacological Interactions of SLITRK1 Protein

Emre Aktaş, Derya Yıldız, Ali Rıza İslim, Nehir Özdemir Özgentürk[1]

[1]Yıldız Technical University

Our study delves into the structural dynamics and pharmacological interactions of the SLITRK1 protein, central to neural development and implicated in neurological disorders such as Tourette

Syndrome. Utilizing advanced bioinformatics tools, we identified the S656 position as crucial for structural disorder, with the S656M alteration significantly impacting protein stability and drug interactions, particularly with FDA-approved neurotherapeutics Pimozide, Aripiprazole, Risperidone, and Haloperidol. Molecular docking and dynamics simulations revealed that the S656M mutation alters binding affinities, suggesting a profound influence on therapeutic efficacy. These insights not only enhance our understanding of SLITRK1's role in neurobiology but also underscore the potential for targeted drug development, highlighting the value of computational methods in elucidating protein interactions and disease mechanisms. Presenting this research at the 2nd ML4NGP Meeting 2024 offers a significant opportunity to contribute to the field, engage with leading experts, and explore the intersection of machine learning and protein study, with the guidance and insights from distinguished figures like Professors Vendruscolo, Orengo, Moult, and Ouzounis being particularly invaluable.

P10

Kinetics and thermodynamics characterization of Serum Amyloid A (SAA) protein

Haidara Nadwa[1], Faidon Brotzakis[2], Annalisa Santucci[1], Daniela Braconi[1]

[1]University of Siena, Italy
[2]University of Cambridge, UK

Serum amyloid A (SAA) protein is an inflammatory acute-phase apolipoprotein that is normally associated with high-density lipoprotein (HDL) in serum and synthesized predominantly by the liver and expressed constitutively or in response to inflammatory stimuli. It is a major component of secondary amyloid deposits which is a systemic amyloidosis ensuing upon long-term complications of several chronic inflammatory disorders. The formation of amyloid fibrils in systemic AA amyloidosis derives from the misfolding of circulating SAA after a significant increment of serum SAA1 level after the dissociation between SAA and HDL. The main treatment approach aims to decrease SAA serum levels, which reduces the amyloid deposits, by removing the underlying inflammatory conditions. However, these available therapies are not always effective due to other factors besides the amyloid fibrils that can contribute to clinical symptoms demanding the development of targeted therapies to inhibit amyloid formation. In this context, to better understand the unfolding mechanism of SAA, it is essential to characterize it in terms of structural and kinetic ensembles to determine its normal and aberrant behavior. To achieve this, we attempted to describe the thermodynamic ensembles that are represented by the free-energy landscape. This would give information about the different metastable states of SAA as well as the respective populations and pathways of interconversion, trying to characterize in detail a possible misfolding pathway. Starting from four folded structures of SAA and four chains of the fibril structure, we run short unbiased molecular dynamics simulations to analyze and construct possible collective variables (CVs) that reveal much of the physics that underlies the folding/unfolding process. Then, using these CVs, we run bias-exchange metadynamics (BE-META) simulations in which the system goes from the folded state to the unfolded one. Eventually, this allows to identify of different metastable states on the proposed transition pathway. Now, we are trying to determine kinetic ensembles of SAA by exploiting the theory of stochastic processes using Markov state models

analyses to gain new insights about the transition rates between the defined metastable states. A better understanding of the structural and kinetic properties of the monomeric state of SAA will be key to determining the effects of small-molecule drug candidates on the aggregation behavior of this protein.

P11

Identification of the NLRP3 structural domain crucial for its activation and new bioinformatics criterion for assessment of mutations as a risk factor for inflammatory diseases

Kristina Stevanović, Vladimir Perović, Sanja Glišić[1]

[1]Vinča Institute of Nuclear Sciences, University of Belgrade, Serbia

An important intracellular sensor in the innate immune system, the NOD-like receptor pyrin domain-containing protein 3 (NLRP3) is capable of detecting a wide range of endogenous and external stimuli, which in turn initiates the creation of inflammasomes and the production of IL-1β. Numerous inflammatory, immunological, viral, metabolic, degenerative, and tumorigenic illnesses have all been connected to its changed activity. Cryopyrin-associated periodic syndrome (CAPS), caused by gain-of-function mutations in the NLRP3 gene, is characterized by the spontaneous formation of inflammasomes in the absence of activating signals. Lately, there has been a growing focus on comprehending the NLRP3 interdomain organization and devising strategies to prevent excessive activity in pathological conditions.
Here we used the informational spectrum technique (ISM) to identify evolutionary highly conserved information encoded in the NLRP3 primary structure responsible for its activation. Our findings lead us to suggest a bioinformatics criterion for estimating the biological impact of nonsynonymous single nucleotide substitution within coding regions of the NLRP3 gene. Additionally, NLRP6, which undergoes liquid-liquid phase separation, was used in comparative bioinformatics analysis. The main distinctions between these two proteins were revealed by the absence of specific peaks in the observed spectra. Molecular dynamics simulations have been used to investigate the interactions of important amino acid residues with several NLRP3 inhibitor candidates that were identified through the use of an enhanced ML screening protocol.

P12

Intra-chain Interactions Shape the Evolution of HSPB1 Chaperone Proteins

Luciana Rodriguez Sawicki, Vanesa Racigh, Maria Silvina Fornasari[1]

[1]Universidad Nacional de Quilmes

Background.
The human proteome encodes ten paralogs of the small heat shock protein family (sHSP) of chaperones, which are overexpressed under stress conditions. They constitute the first line of defense in the heat shock chaperoning network, interacting transiently in an ATP-independent

manner with client proteins to prevent their misfolding and aggregation. From a structural standpoint, sHSPs are composed of a structured domain known as the alpha crystallin domain (ACD), flanked by the N- and C-terminal regions (NTR and CTR), which exhibit variable lengths and are intrinsically disordered.
Recent experiments have demonstrated that the end of the NTR can establish reversible contacts with a pocket located within the ACD. These cis-interactions have been proposed as a regulatory mechanism for the chaperone activity of small heat shock protein 1 (HSPB1). Considering that inter-domain interactions impose constraints on amino acid substitutions during evolution, the objective of this study is to investigate the impact of these interactions on the evolution of HSPB1.
Method and results
We retrived orthologous sequences of Homo sapiens HSPB1 from the protein database at NCBI. The dataset underwent manual curation to remove low-quality and incomplete sequences. Multiple sequence alignment was performed using Clustal Omega, and evolutionary rates per position were calculated using the Rate4Site software. Phylogenetic trees were constructed using the PHYML program with the maximum likelihood method and bootstrap support. Additionally, we performed Direct Coupling Analysis (DCA) to study residue co-evolution in HSPB1. By using a sequence alignment of a newly curated dataset of homologous sequences recruited using blastp, the couplings were calculated using the mean field algorithm. The scores ranking each pair of interacting sites were computed using the direct information method and the top contacts were mapped onto the structure of HSPB1. Preliminary results indicate that the rate of amino acid evolution in the terminal regions differs from that of the ACD, consistent with existing literature. However, there are subregions of the NTR and CTR showing a low rate of evolution. Furthermore, amino acids belonging to these regions are implicated in contacts with the ACD. Results from the DCA method are also in agreement with experimental results reported in literature.
Conclusions
These findings suggest that the evolution of the NTR and ACD is not independent; instead, these regions have co-evolved to maintain the function and regulation of HSPB1.

P13

Protein Function Prediction: A Graph-Based Approach

Mahta Mehdiabadi

University of Padova, Italy

In this study, we propose a novel approach to tackle the protein function prediction problem using Graph Neural Networks (GNNs). Understanding protein functions is crucial for deciphering cellular processes and developing targeted therapeutics. Our method leverages the inherent graph structure of domain and disorder region interactions, capturing intricate relationships among different domains and features.
We employ GNNs to learn informative domain representations, effectively encoding the complex dependencies within the domain interaction graph. Our model integrates multi-modal data, including sequence information, structural information, and functional annotations, providing a comprehensive framework for prediction. Preliminary evaluations on benchmark datasets

demonstrate promising results, indicating the potential effectiveness of our approach compared to traditional methods.
This research contributes to the evolving landscape of protein function prediction, presenting a robust and interpretable tool for biologists and bioinformaticians. The integration of graph neural networks opens new avenues for uncovering intricate protein relationships, fostering a more nuanced understanding of Gene Ontology annotations.

P14

Self-assembling protein-based nanoparticles for biomedical and biotechnological application

Marc Fornt-Suñé[1], Gonzalo Lázaro Bermejo[1], Marcos Gil-Garcia[1], Andrea Aran[1], Javier Garcia-Pardo[1]*, Mercè Martí[1]*, Salvador Ventura[1]*

[1]Institut de Biotecnologia i de Biomedicina (IBB), Universitat Autònoma de Barcelona, Spain

Coiled coils are characterized by an arrangement of two or more α-helices into a superhelix, and they are abundant structural motifs in nature as they are found in many fibrous proteins and transcription factors. Additionally, some coiled coils have self-assembling properties that allow them to interact with each other and form higher-order protein structures. Hence, this fold is of significant interest for rational protein design as it can be used as a building block to create novel protein-based nanostructures. We exploited the self-assembling properties of ZapB, a natural anti-parallel coiled coil-forming prokaryotic protein, to build α-helical sustained nanostructures that can be readily purified from microbial cell factories in a pre-assembled state. The strategy relies on the creation and expression of a fusion protein, which consists of ZapB fused to different functional domains that will be embedded in the resulting coiled coil-based self-assembling nanoparticles. This modular approach allowed us to create two distinct sets of functional protein nanoparticles by simply changing the functional domains that are sequentially fused to ZapB. The first set exhibits high affinity antibody-binding properties and can be easily decorated with any combination of antibodies in the desired ratio. In a proof-of-concept study, these protein nanoparticles were decorated with anti-human CD3 and anti-human CD28 antibodies and were implemented to stimulate the activation and proliferation of human T cells, emerging as a promising material with potential application in cancer immunotherapy. In contrast, the second set of protein nanoparticles are able to selective capture SARS-CoV-2 virus, presenting a promising avenue for the development of new potent antiviral materials and virus-filtering nanodevices. In light of the above, this versatile protein design strategy opens the door to the development of novel functional protein-based nanoparticles for different biomedical and biotechnological applications.

P15

Exploration of the vibrational space of tandem repeat proteins

Matias Chiappinelli, Tadeo E. Saldaño, Sebastian Fernandez-Alberti[1]

[1]Universidad Nacional de Quilmes, Departamento de Ciencia y Tecnología, Roque Saenz Peña 352, Bernal, Argentina.

Protein tandem repeats are composed of adjacent repetitive units with similar sequence motifs. Their biological function can be enlightened by differential sequence, structure and dynamics aspects provided by their tandem repetitive motifs with respect to those corresponding to their individual units. In this work, we explore their dynamics aspects by using Normal-modes analysis (NMA) based on a coarse-grained elastic network atom contact model particularly developed to simulate protein dynamics and the prediction of the effect of mutations. As a test case, we have considered the protein phosphatase 2A (PP2A) that presents an alpha-solenoid structure, characterized by solenoidal windings of the polypeptide chain having exclusively alpha-helices as the secondary structure elements.
We explore effects of mutations on different aspects of the dynamics of the 15 tandem repeat units of the A subunit of PP2A and relate them with position-specific evolutionary conservations obtained from multiple sequence-structure alignments. For this purpose, the effect of mutations on different selected sets of vibrations were simulated. On one hand, we analyze subspaces of vibrations related to collective dynamics, i.e., dynamics involving several repetitive units that are commonly related to the protein scaffold. On the other hand, we carried out comparative studies of the dynamics of different individual units according to their symmetric location in the global protein. Furthermore, we explore correlated motions between units and how they impact on the evolution of their sequence profiles.

P16

Machine-learning for data-driven protein engineering

Mehdi Davari

Leibniz Institute of Plant Biochemistry, Germany

Protein engineering is nowadays used routinely in developing biocatalysts for biotechnological, biomedicine, and life sciences applications. Recently, data-driven strategies attracted attention in protein design and engineering because of advances in large experimental databanks of proteins, next-generation sequencing (NGS), high-throughput screening (HTS) methods, and the development of artificial intelligence algorithms [1-3]. However, the reliable prediction of beneficial amino acid substitutions, their combination, and effect on functional properties still pose significant challenges in machine learning assisted protein engineering [4]. In this presentation, we describe a general-purpose framework (PyPEF: Pythonic Protein Engineering Framework[5-6]) for data-driven protein engineering by combination of machine learning methods with signal processing and statistical physics techniques. PyPEF assist in the identification and selection of beneficial proteins of a given sequence space by either systematically or randomly exploring the fitness of variants and by sampling random evolution pathways. The predictive accuracy and throughput performance of PyPEF was evaluated based on four public protein and enzyme datasets using common regression models. It turns out that the PyPEF could efficiently predict the fitness of protein sequences for different target properties (predictive models with coefficient of determination values ranging from 0.58 to 0.92). By combining machine learning and protein evolution, PyPEF enabled the screening of

proteins with various functions reaching a screening capacity of more than 500,000 protein sequence variants in the timeframe of only a few minutes on a standard PC. PyPEF exhibited significant accuracies on four public datasets (different proteins and properties) and underlined the potential of integrating data-driven technologies for covering different philosophies by either predicting the fitness of the variants to the highest accuracy or capturing the general trend of introduced mutations on the fitness in directed protein evolution campaigns. In essence, PyPEF provide a powerful solution to current sequence exploration and combinatorial problems present in protein engineering through exhaustive in silico screening of the protein sequence space.

P17

Exploring the Impact of Intrinsically Disordered Regions on RNA Stability and microRNA Accessibility

Mert Cihan, Gustavo Parisi, Miguel A Andrade-Navarro

Intrinsically disordered regions (IDRs) in proteins are well-recognized for their critical roles in mediating interactions with DNA and proteins, attributed to their flexibility and among others, the presence of repetitive amino acid sequences. Beyond these established roles, this study explores the influence of IDRs at the transcript level, particularly on RNA stability and microRNA binding site accessibility. This investigation is motivated by the observation that the characteristics of IDRs, such as repetitive sequences that confer dynamic functionality to proteins, may similarly enhance RNA stability, thereby altering the landscape for microRNA interaction.
MicroRNAs, short non-coding RNAs that regulate gene expression within the RNA-induced silencing protein complex (RISC), require structural accessibility to their target sites in the 3' untranslated regions of messenger RNAs. This interaction typically results in the downregulation of gene expression, either through the degradation of mRNA or the repression of its translation, effectively modulating protein synthesis in the cell. We hypothesize that increased RNA stability in transcripts coding for IDPs (intrinsically disordered proteins) could deter microRNA binding and circumvent translational repression, thus revealing a novel aspect of IDP functionality at the RNA level.
To test these hypotheses, we mapped consensus IDRs across protein transcripts, evaluating both global and local RNA secondary structures based on minimum free energy principles. We also investigated the effect of splicing on the presence of IDRs and its correlation with RNA stability, in addition to analyzing the frequency of secondary structural features, like hairpin loops, within microRNA-targeted regions. Our findings indicate a higher stability in RNA transcripts of IDPs compared to those without IDRs, due to alternative splicing events, with notable stability in transcripts that contain conserved microRNA binding sites for specific microRNA families. This stability is characterized by the prevalence of structural features conducive to stability, such as hairpin loops, unlike the more disrupted structures seen in non-IDP transcripts having binding sites for the same microRNA families, which exhibit internal loops with consecutive mismatches.
In summary, our study highlights the crucial role of IDRs in mediating not only protein-DNA interactions but also in affecting RNA stability and microRNA accessibility. This expanded functionality of IDPs suggests a complex molecular interplay with significant implications for our understanding of gene regulation mechanisms.

P18

Structural determinants of a-synuclein binding to an inhibitory peptide studied by molecular dynamics simulations

Oriol Bárcenas[1,2], Salvador Ventura[1], Ramon Crehuet[2]

[1]Autonomous University of Barcelona (UAB), Spain
[2]CSIC-Institute for Advanced Chemistry of Catalonia (IQAC), Spain

Parkinson's disease is the world's second most prevalent neurodegenerative disease and is characterized by the loss of dopaminergic neurons in the substantia nigra pars compacta. a-Synuclein aggregation into amyloid fibers is the main pathological hallmark of Parkinson's disease. Thus, it is the focus of many studies that aim to understand and find treatment for this debilitating disorder.
a-Synuclein is an Intrinsically Disordered Protein: it does not have a defined structure in solution. Thus, it cannot be studied using classical methods that rely on a single structure, such as X-ray crystallography or Cryogenic Electron Microscopy (CryoEM). On the other hand, methods that measure the protein's structure in solution, such as nuclear magnetic resonance (NMR) or small-angle X-ray scattering (SAXS), are more adequate but only produce ensemble-averaged data with reduced information content. These methods provide only marginal information on a-Synuclein conformation, and alternative strategies are required, such as computational models and molecular simulations.
Given the importance of Parkinson's disease, many efforts have focused on finding therapeutics targeting the formation of aggregates. In one such study, PSMa-3, an antimicrobial peptide synthesized by S. aureus, was identified as a highly potent inhibitor of the aggregation of a-Synuclein. It exerts this activity by binding preferentially to an a-Synuclein region comprising residues 24-58, which is highly relevant for Parkinson's disease, as almost all inherited mutations of familial Parkinson's disease are located in this region. For this reason, the peptide of the 24-60 amino acid region is simulated.
This work focuses on simulating and studying the interaction of a-Synuclein with PSMa-3 in solution and elucidate whether its interaction significantly impacts their structure. Each peptide will also be analyzed separately to compare the obtained results. To this end, the Molecular Dynamics open-source software GROMACS and supercomputing resources are used. Enhanced sampling using ""Parallel Tempering, Well-Tempered Ensemble"" (PT-WTE) is employed to visit a more expansive conformation space of both peptides and their complex. With this method, rare conformations are visited more often than would otherwise be feasible with a classical simulation method.
The simulation results agree with the experimentally observed behavior of the two peptides. Two distinct interaction surfaces between PSMa3 and a-Synuclein are described, with strand secondary structure governing these contacts. Moreover, these contacts can induce secondary structures (strand and helix) in neighboring residues. Finally, the key role of the hydrophobic effect in promoting these contacts is described.

P19

Deciphering Protein-Peptide Interactions: Unveiling the Intricacies of PRM Binding to SH3 Domains

Radovan Dvorsky, Ondrej Cehlár[1]

[1]Institute of Neuroimmunology, SAS, Bratislava, Slovakia

Interaction of proteins with peptides is one of the most important processes in biology and fundamental in the transmission of signals within the cells. Crucial for such interaction is the recognition of peptides sharing some characteristic features by globular proteins or globular protein domains that typically also share structural fold and usually belong to one protein family. Complexes between such proteins and peptides can be viewed as switch hubs in the cell signaling networks. An excellent example of such protein-peptide system are Src Homology domains on the side of protein and peptide fragments containing prolines. Former ones, known as SH3 domains are small protein modules play crucial role in cell growth, differentiation, and apoptosis, later known as proline-rich motifs (PRMs) are generally unstructured. Because PRMs do not adopt a defined secondary structure, their binding to SH3 domains often relies on induced fit mechanisms, where the PRM undergoes conformational changes upon binding. However, SH3 domains can recognize different PRMs through distinct mechanisms providing thus a way for very versatile networking of proteins. How the recognition is achieved and what determines that some PRMs are recognized while other do not associate at all is the main goal in the research of this system. In an attempt to solve this, we have studied and analyzed several interactions between particular SH3 domains and PRMs experimentally as well as collected sequential, structural information in a genome wide context. Although this system can be studied by some classical approaches utilizing the methods of molecular modelling or related techniques we have set out to tackle this problem by machine learning. In the poster I will describe the whole background, represent our recent achievements as well as drawbacks with the aim to discuss possible future ways of solutions.

P20

Analysis of tandem repeat containing proteins in Alpha- and Betabaculovirus

Santiago Manuel Gomez Bergna[1], Miguel Andrade-Navarro[2], Maria Leticia Ferrelli[1]

[1]Instituto de Biotecnología y Biología Molecular - UNLP-CONICET, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, La Plata 1900, Buenos Aires, Argentina
[2]Insitute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University, Biozentrum I, Hans-Dieter-Hüsch-Weg 15, 55128 Mainz, Germany

Baculoviridae is an insect-specific virus family and is subject of extensive research mostly in pest biocontrol and biotechnology. Alpha- and Betabaculovirus are two baculovirus genera that infect insects of the order Lepidoptera. They have circular dsDNA genomes of approximately 80-180k bp length and around 150 coding genes.

Tandem repeat-containing proteins are those that feature amino acid motifs repeated two or more times and located adjacent to each other. It's known that TRs of proteins are involved in protein-protein interactions, signal transduction, DNA/RNA binding, and structural stability, among others. It's also known that TR regions have very specific evolutionary dynamics.
The aim of this work is to study the distribution of TR regions across the proteomes of these two baculoviral genera deposited in UniProt. To perform this, we downloaded the proteomes and used the Meta-Repeat-Finder (MRF) pipeline to predict TR regions in each genera. In parallel, we used OrthoFinder to cluster sequences of alpha and betabaculoviral proteomes to group them in different orthogroups. Finally, we merged this information to analyze the prediction of TR regions in each dataset.
After running MRF against alphabaculovirus and betabaculovirus proteomes we obtained 3457 and 2013 TR regions for each genera, respectively. When analyzing orthogroups within these datasets, we found 911 orthogroups containing at least two proteins: 499 containing only alphabaculoviral proteins, 261 containing only betabaculovirus proteins, and 152 containing proteins of both genera. While analyzing the presence of TR in orthogroups, we found 419 alphabaculoviral orthogroups with TR regions, 201 betabaculoviral orthogroups with TR regions, and 70 with TR regions belonging to both.
Separating TRs based on the repeat unit length, we found that the most frequent TR group among alphabaculoviral proteins were those with repeat length greater than 50 aminoacids, while the most frequent among betabaculoviral proteins were TRs with repeat length between 9 and 49 aminoacids. An interesting result is that although TR regions were predicted in many proteins of the same orthogroup, they were not predicted in all members. In ongoing work we are evaluating whether these differences reflect evolutionary and functional variation.
In conclusion, we could identify several orthogroups, containing or not TR regions. We also identified that even if some proteins are grouped in the same orthogroup, they might not contain the same TR regions. These results will allow us to better understand the evolution and functionality of TR regions in baculoviral species.

P21

Investigating conformational space of protein tau by enhanced sampling MD simulations

Stefana Njemoga[1,2], Vojtech Spiwok[3], Ondrej Cehlar[1]

[1]Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia
[2]Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia
[3]Department of Biochemistry and Microbiology, University of Chemistry and Technology, Prague, Czech Republic

Turning IDPs into druggable targets provides a great opportunity to extend the druggable target-space for novel drug discovery. Since, the most accurate way to describe IDPs is by providing their conformational ensemble, ensemble-based approach involving generation of realistic ensemble of conformers subjected to binding sites analysis is generally chosen as a drug design

strategy. To this end, ensemble generators or molecular dynamics simulation can be used to produce the initial set of protein conformers. A common undesirable result of MD simulation is conformation trapping in specific energetic minimum, leading to insufficient conformation sampling. Biasing MD simulations by altering the potential energy function can help protein to overcome energetic barrier and more efficiently explore its conformational space. With this aim, we employed metadynamics simulations with two sets of collective variables (CVs), with the first set being CVs derived from alphafold, and second set of CVs describing the protein's secondary structural properties by stimulation of alpha-helix and anti-beta structure formation. In addition, third set of CVs has been designed by t-distributed parametric time-lagged stochastic neighbor embedding (ptltSNE) machine learning method for future simulations. By comparing results from different types of enhanced sampling simulations, we can select the most suitable method for conformational study of non-globular protein tau and obtain new metastable conformations of tau for our future experiments. The ultimate goal of the study is generation of realistic conformational ensemble of protein tau in order to identify potential binding sites targetable by small molecules.

P22

TAC-1: Advancing towards the development of small molecules to treat tauopathies

Zoe Manglano-Artuñedo[1], Samuel Peña-Díaz[1], Jaime Santos[1], Helena Østergaard Rasmussen[2], Jan Skov Pedersen[2], Daniel E Otzen[2], Irantzu Pallarès[1], Salvador Ventura[1]

[1]Institute of Biotechnology and Biomedicine, Autonomous University of Barcelona, Barcelona, Spain
[2]Aarhus University, Aarhus, Denmark

Tauopathies are a set of neurodegenerative diseases characterized by the amyloid aggregation of an intrinsically disordered protein named tau. The most common tauopathy is Alzheimer's disease, which is the main cause of dementia worldwide and leads to a high social and economic burden on our society. Nowadays, the current therapies are only addressed to treat the symptomatology, but not the onset nor progression of these disorders. Here, we developed a novel high-throughput screening methodology to find new tau aggregation inhibitors based on the implementation and optimisation of seeded tauK18 aggregation kinetics. We obtained 97 compounds with anti-aggregating properties that subsequently were tested on aggregation kinetics with full-length tau (Tau2N4R). This two-step screening process (TauK18 and Tau2N4R) allowed us to identify TAC-1, a promising compound that inhibits the amyloid aggregation of the two variants as reported by thioflavin-T and light scattering. We found that TAC-1 is an inhibitor with concentration-dependent activity that may impact aggregate structures. Flow Induced Dispersion Analysis and Small Angle X-Ray Scattering results suggest that TAC-1 targets Th-T binding species altering the structure of tau aggregates. Taken together, we conclude that TAC-1 may constitute a promising hit compound to inhibit and modulate tau amyloid aggregation.

P23

Combined in silico and in tube tools to study the structure of intrinsically disordered tau proteins

Adam Polák, Stefana Njemoga, Katarína Martonová, Katarína Tomková, Ondrej Cehlár, Rostislav Škrabana[1]

[1]Institute of Neuroimmunology, Slovak Academy of Sciences, Bratislava, Slovakia

The lack of therapeutics to treat Alzheimer's disease (AD) and its increasing worldwide prevalence are showcasing the shortage of insight into the pathogenesis of the disease. The main topic of my doctoral project is to analyse small structural motifs potentially involved in tau protein aggregation in AD. Tau protein is responsible for the formation of insoluble neurofibrillary tangles, what are one of the main hallmarks of the AD. The study of tau structure is limited by the fact that tau belongs to the family of intrinsically disordered proteins (IDPs). Moreover, tau protein is disordered almost in its full length and its conformation is depending on the binding with its binding partners.
As the crystallization of IDPs is unachievable with standard crystallographic approaches, we utilise co-crystallization of tau protein with monoclonal antibodies, specifically its fragment antigen-binding regions (Fabs). Fabs are able to stabilize tau protein in a transient complex and enable to characterise its epitopes and structural motifs.
Machine learning tools are another useful approach that can be used to predict otherwise unsolvable structure of IDPs as well as their binding-sites with potential interaction partners. Our aim is to use the predictor software such as Alphafold multimer or Cabsdock to investigate possible formation of structural motifs on free tau, inspired by similar tau conformation in complexes with monoclonal antibodies. After successful in silico part of the project, the interacting site and the structural motifs can be mutated to determine their role in tau aggregation.
Our next aim is to prepare complexes of three monoclonal antibodies, MN423, DC11 and DC25 with tau 297-391, obtain crystals of these complexes and proceed with diffraction analysis to identify larger tau protein structure and its role in tau aggregation. Results obtained from in silico and experimental structural analysis of tau protein are applicable in the design of AD targeting therapeutics and the development of AD-specific vaccine, both leading to improved management of the disease.
This project is supported by APVV-21-0479, VEGA 2/0141/23, 2/0125/23 and EU Grants ADDIT-CE, InterTau.

P24

Tetrahydrofurans derivatives inhibit Aβ amyloid fibrillation

Andrea Antosova[1], Zuzana Bednarikova[1], Angelika Lasikova[2], Jana Novacikova[3], Tibor Gracza[2], Zuzana Gazova[1]

[1]Department of Biophysics, Institute of Experimental Physics Slovak Academy of Sciences, Kosice, Slovakia

[2]Department of Organic Chemistry, Faculty of Chemical and Food Technology, Slovak University of Technology, Bratislava, Slovakia
[3]Department of NMR Spectroscopy and Mass Spectrometry, Faculty of Chemical and Food Technology, Slovak University of Technology, Bratislava, Slovakia

Amyloid β peptide (Aβ) is the main component that plays a significant role in the pathogenesis of Alzheimer's disease (AD) and is considered to be the leading cause of AD development. Diverse strategies have been employed to identify Aβ peptide amyloid aggregation inhibitors, including small organic molecules as furan, curcumin, quercetin, polyphenols and many other synthetic and natural compounds. Many effective small molecule inhibitors are planar molecules bearing a lipophilic bi- or tri-(hetero)-cyclic scaffold and having basic nitrogen on one of the rings or having a polyphenolic structure.
We synthesized a set of alkenols (W10-W14) in the construction of (poly)hydroxylated tetrahydrofurans (THF). We investigated their effect on Aβ1-40 amyloid fibril formation using ThT fluorescence assay and atomic force microscopy (AFM). Compounds W10, W11, and W14 were optically pure isomers, and compounds W12 and W13 were diastereomeric mixtures of 2,3,4-trisubstituted THF.
Our findings indicate that the studied compounds exhibit differential inhibitory effects on Aβ1-40 amyloid formation. Optically pure derivatives W10, W11 and W14 with substituents at C5 showed high inhibiting activity (~70-90%). No inhibitory effect on amyloid fibrillization was observed for THF derivatives W12 and W13. The morphology and amount of amyloid aggregates formed in the presence of THF were determined by AFM and support data obtained by ThT assay. The obtained data suggest a correlation between the chemical structure of studied THF derivatives and inhibitory activities in in vitro experiments. We assume that the presence of aromatic moiety in the C5 position and the configuration of the THF derivative determine their inhibiting activity. The observed results suggest that THF compounds W10 and W11, characterized the best inhibitory activity in the μM range, can potentially be therapeutical agents for diseases associated with amyloid aggregation of the Aβ1-40 peptide.
This work was supported by the Slovak Research and Development Agency under Contract No. APVV-22-0598 and Slovak Grant Agency VEGA 2/0176/21, 02/0164/22. We are thankfull to Katarina Siposova for Thioflavin T data acquisition.

P25

Effects of the ionic strength on functional amyloid fibril formation

Andrea Bartolomé-Nafría[1], F. Emil Thomasen[2], Kresten Lindorff-Larsen[2], Salvador Ventura[1], Javier Garcia-Pardo[1]

[1]Universitat Autònoma de Barcelona, Spain
[2]Københavns Universitet

Protein aggregation depends on two main factors: the primary sequence of the protein and the environmental conditions. Therefore, for the same protein sequence, changes in pH, temperature or solvent composition can influence the aggregation kinetics. In a similar manner, the presence of

metal ions is thought to play a crucial role in many biological functions, including protein folding, stability and amyloid aggregation. In this context, the study of the environmental effects on amyloid aggregation can provide insights into the differences in disease progression and the in vitro assembly of amyloids for technological purposes.
hnRNPDL is a ribonucleoprotein (RNP) involved in transcription and RNA processing that harbors missense mutations causing limb-girdle muscular dystrophy D3 (LGMD D3). Mammalian-specific alternative splicing (AS) gives rise to three isoforms, hnRNPDL-2 being the predominant in human tissues and the only isoform capable of rendering functional amyloid fibrils with nucleic acid-binding properties, suggesting that LGMD D3 might be a loss-of-function disease associated with impaired fibrillation.
In the present work we have investigated how solvent ions influence hnRNPDL-2 amyloid aggregation. By using SAXS and computational simulations, we have explored the changes in hnRNPDL-2 structure induced by the presence of different metal ions and salt concentrations. In vitro aggregation experiments and structural data support our findings, suggesting that the ionic strength significantly affects hnRNPDL-2 self-assembly properties. Ultimately, we believe our findings can help understanding disease cause and progression associated with this protein and other functional amyloids involved in human diseases.

P26

Analysis of the impact of disordered insertions in ensembles of structural tandem repeats

Carla Padilla Franzotti[1], Nicolas Palopoli[1], Gustavo Pierdominici-Sottile[1], Miguel Andrade[2]

[1]Department of Science and Technology, National University of Quilmes, Buenos Aires, Argentina
[2]Computational Biology and Data Mining Research Group, Johannes-Gutenberg University, Mainz, Germany

Structured tandem repeats (STRs) play crucial roles in protein function and evolution. Insertions within ensembles of STRs introduce diversity and potentially alter their properties. We study the functional implications of insertions in STR ensembles using a multifaceted approach:
Amino acid composition analysis: characterizing the inserted sequences to identify unique compositional features.
Predicted structures: using computational tools to assess the impact of insertions on the folding and stability of the STR ensemble.
Conservation analysis: leveraging the ELM database to identify conserved residues within insertions and potential functional motifs.
Dynamic simulations: employing simulations to explore the conformational dynamics of STRs with and without insertions.
By integrating these diverse methods, we aim to gain a comprehensive understanding of how insertions modulate the properties of STR ensembles, ultimately gaining information on their functional significance and evolution.

P27

Systematic Comparative Functional Analysis of Protein Regions Identified by Complementary Approaches to Delineating Compositional Bias, Structural Disorder, and Low Complexity

A Gruca[1], D P Kreil[2]

[1]Silesian Univeresity of Technology
[2]Boku University Vienna, Austria

The biological importance and the functional roles and of compositionally biased, structurally disordered, and low-complexity sequence regions are increasingly recognized. Most reports, however, remain anecdotal or focus on the characterization of a particular method for delineating a region of interest due to its unusually low variance in amino acid composition, lack of structural order, or low information complexity. We here present a recent study that, for the first time, systematically compares functional associations of proteins featuring regions identified by a range of complementary approaches to delineating compositional bias, structural disorder, and low information complexity in amino acid sequences. Care is taken to remove artefacts from unbalanced ontological representations in the knowledge base, propagating knowledge along the hierarchical structure. Rigorous statistical significance is provided through resampling in line with the population density of that representation. For a layered interpretation, functional enrichment is assessed relative to negative sets matched for protein sequence characteristics. We extend analyses comprising complementary functional ontologies such as the GeneOntology, KEGG, etc. with systematic queries of the scientific literature relevant to the identified proteins, exploiting independently validated and benchmarked large language models. Future work will consider positional effects of the identified regions relative to the overall protein sequence.

P28

The role of S100A1, S100A8 and S100A9 heterodimerization in modulating their amyloid aggregation

Darius Šulskis, Ieva Baronaitė, Viktorija Karalkevičiūtė, Vytautas Smirnovas[1]

[1]Sector of amyloid research, Institute of Biotechnology, Vilnius University, Vilnius, Lithuania

Background: The S100A family consists of calcium binding proteins that are largely known for their contribution to the inflammatory processes [1]. Most of the S100A proteins are found as homodimers, but a select few can form heterodimers, that alter their function and structure. Among these S100A8/S100A9 dimer is the most investigated [2], but other members are also capable of forming heterodimers. Furthermore, S100A proteins can assemble into larger oligomers and even amyloid fibrils [3], a hallmark of neurodegenerative disorders. In this study, we aimed to elucidate the aggregation propensities of S100A8, S100A9, S100A1, and their heterodimerization by investigating aggregation kinetics, secondary structures and morphologies of aggregates.
Methodology: Aggregation was monitored with amyloid-specific dyes (ThT, ThS). Fourier-transform infrared (FTIR) and circular dichroism (CD) spectroscopy were used to determine the secondary

structure. The morphology of fibrils was observed with fluorescence, atomic force and electron microscopies.
Results: For the first time, we observed in vitro aggregation of S100A8, which formed spherical aggregates, unlike fibrillar structures of S100A9 under the same conditions. The formation of S100A8 and S100A9 heterodimer stabilized the proteins in the absence of calcium ions and prevented their amyloid aggregation. On the other hand, S100A1 had minimal impact on S100A9 aggregation. but the interaction between S100A1 and S100A8 was modulated by calcium ions. Amyloid fibril formation was observed at equimolar protein to calcium concentrations with higher calcium concentrations inhibiting aggregation. These findings indicate specific correlations between S100A proteins dimerization, stability and amyloid formation.
Conclusion: Formation of S100A8/S100A9 heterodimer inhibited amyloid aggregation, whereas S100A1/S100A8 dimer aggregation was modulated by calcium ions.

References:
[1] L. Santamaria-Kisiel, A. C. Rintala-Dempsey, and G. S. Shaw, 'Calcium-dependent and -independent interactions of the S100 protein family.', Biochem. J., vol. 396, no. 2, pp. 201–214, 2006, doi: 10.1042/BJ20060195.
[2] T. Vogl, N. Leukert, K. Barczyk, K. Strupat, and J. Roth, 'Biophysical characterization of S100A8 and S100A9 in the absence and presence of bivalent cations', Biochim. Biophys. Acta - Mol. Cell Res., vol. 1763, no. 11, pp. 1298–1306, 2006, doi: 10.1016/j.bbamcr.2006.08.028.
[3] S. B. Carvalho et al., Structural Heterogeneity and Bioimaging of S100 Amyloid Assemblies. Elsevier, 2013. doi: 10.1016/B978-0-12-394431-3.00018-3.

P29

Chemically-driven dynamics of disordered proteins and their condensates in simulations

Emanuele Zippo[1], Friederike Schmid[1], Thomas Speck[2], Lukas Stelzl[1,3]

[1]JGU Mainz
[2]University of Stuttgart
[3]IMB

Understanding the condensation and aggregation of intrinsically disordered proteins (IDPs) in a non-equilibrium environment is crucial for unraveling many biological mechanism.
We can now address this with residue-level coarse-grained Molecular Dynamics simulations, integrating Metropolis Monte Carlo steps to model chemical reactions.
We investigate TDP-43 phosphorylation by CK1d enzyme in simulations, examining patterns of phosphorylation and assessing its preventive role in chain aggregation, possibly associated with neurodegenerative diseases.
We find that the degree of residue phosphorylation is determined by sequence preference and charges, rather than the position in the chain.
Depending on the sequence context, phosphorylation stabilizes or de-stabilizes condensates.

For TDP-43, our simulations show condensates dissolution through phosphorylation, in accordance with experiments.
The disordered tail of the kinase Ck1d drives recruitment to the condensates.
To further explore the dynamics of non-equilibrium steady-state systems, like our target system, we apply Markov state modelling. We use VAMPnet neural network to coarse-grain trajectories into Markov states, with which we are able to verify the thermodynamic consistency of the phosphorylation step.

P30

Training and collaboration opportunities in DisProt, the database of Intrinsically Disordered Proteins

Federica Quaglia[1], Maria Cristina Aspromonte[2], Maria Victoria Nugnes[2], Silvio C.E. Tosatto[2]

[1]Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies, National Research Council (CNR-IBIOM), Bari, Italy
[2]Department of Biomedical Sciences, University of Padova, Padova, Italy

The DisProt database (URL: https://disprot.org/) plays a crucial role in the study of intrinsically disordered proteins (IDPs), proteins characterized by the lack of stable three-dimensional structures and involved in a plethora of biological functions. With the introduction of thematic datasets in December 2020, DisProt has enriched its offerings, focusing on key biological processes and organisms through expert collaborations. These datasets, targeting areas like cancer, autophagy, and neurodevelopmental disorders, enhance DisProt's relevance for specialized research, promoting a more organized and collaborative approach to protein science.
To support the ongoing enrichment and growth of DisProt, comprehensive training programs have been developed for new and existing biocurators. These include a flexible e-learning course, offered in both English and Spanish via the ELIXIR-SI eLearning platform, accessible year-round. Complementing these are live Zoom sessions and in-person training events that provide hands-on experience and opportunities for direct interaction with experts in the field. These sessions are designed to immerse participants in the practical aspects of biocuration and the use of DisProt, alongside MobiDB, for exploring IDPs. Specialized courses also cover experimental techniques pivotal to the study of IDPs, such as NMR, SAXS, FRET, and CD, guiding researchers on how to interpret and curate data related to disordered proteins or regions.
The emphasis on training and collaboration in DisProt not only underscores the importance of community engagement in the biocuration process but also highlights the database's role in advancing the understanding of IDPs. By fostering a collaborative environment and providing resources for targeted curation and research, DisProt is a valuable and accessible resource for researchers worldwide, promoting collaboration and enhancing the knowledge of intrinsically disordered proteins.

P31

Analysis and Sampling of Molecular Simulation with extended Autoencoder architectures

Guglielmo Tedeschi[1], Vladimír Višňovský[2], Aleš Křenek[2], Vojtech Spiwok[1]

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

Nowadays molecular dynamics (MD) has become a standard tool to investigate molecular systems. By simulating the dynamic behavior of molecules, it facilitates the exploration of stability, conformational changes and a plenty of other properties essential for understanding molecular functions. However, the application of MD is affected by the large computational costs. It computes steps that must be in order of femtoseconds, to assure numerical stability, for a time scale long enough to likely catch some or rarely occurred processes. There are numerous techniques to enhance simulations to optimize the sampling of the slow motions, one of those is metadynamics. It operates by biasing the Hamiltonian of the system, encouraging it to cross barriers and explore configurations that might otherwise be inaccessible or challenging to sample, in a reasonable amount of time. However, to make metadynamics successful, the scientist has to select the so-called “collective variables" which are functions of Cartesian coordinates. Designing good collective variables is not a trivial task and it relies on the knowledge of the system and experience of the scientist. Machine learning and artificial neural networks have shown incredible power in supporting the exploration of conformational phase space.
Hereby we show the involvement of a method based on the control and optimization of the latent space. The developed neural network exhibits the ability to analyze simulation data and to derive optimal combinations of internal coordinates to be used as CVs. The power and the efficiency of the presented approach are demonstrated on Trp-cage foldings and Villin-headpiece.

P32

DPEnsembleTools, A Python package for a comprehensive analysis of conformational ensemble of intrinsically disordered protein

Hamidreza Ghafouri[1], Giacomo Janson[2], Silvio Tosatto[1], Alexander Miguel Monzon[3]

[1]Department of Biomedical Sciences, University of Padova
[2]Department of Biochemistry and Molecular Biology, Michigan State University
[3]Department of Information Engineering, University of Padova

Intrinsically Disordered Proteins/Regions (IDPs/IDRs) constitute over thirty percent of the human proteome, playing pivotal roles in cellular functions like cell signaling, molecular complex formation, and transcriptional regulation. Unlike the singular structural entities which are defining globular proteins, IDP/Rs exist as dynamic ensembles, comprising rapidly interconverting conformations.

Therefore, to unravel the molecular underpinnings of IDPs/IDRs functions, understanding the ensemble properties is imperative.
Introducing DPEnsembleTools (Disordered Protein EnsembleTools), an intuitive Python package designed to decode the conformational ensemble of disordered proteins using structural data. This versatile tool offers a wealth of insights into the ensemble's characteristics, including distributions of phi/psi angles, radius of gyration, asphericity, internal distance distribution, and end-to-end distance (Re), among others. Moreover, DPEnsembleTools facilitates dimension reduction techniques such as Principle Component Analysis, t-SNE, and Force Scheme, enabling effortless representation of ensemble features in two or three dimensions. This capability is instrumental in discerning differences among ensembles derived from various experimental methodologies.
Illustrating its utility, DPEnsembleTools is applied to multiple entries in the Protein Ensemble Database (PED), showcasing its efficacy in analyzing and visualizing complex ensemble data. By empowering researchers with comprehensive ensemble information and advanced visualization techniques, DPEnsembleTools facilitates a deeper understanding of the functional dynamics of IDPs/IDRs.

P33

Revealing Novel Protein-Protein Interaction Modulators of Pyruvate Kinase M2 (PKM2) Enzyme using Structure-Based Drug Design (SBDD)

Ian Czarnowski[1], Romina Armando[1], Patricio Chinestrad[1], Alejandra Zinni[1], Pablo Lorenzano Menna[1], Diego Mengual Gomez[2], Maia Cabrera[1]

[1]Molecular Pharmacology Lab - Science and technology Department - UNQ
[2]Molecular Oncology Unit - Molecular and Translational Oncology Center – UNQ

The aerobic glycolysis (i.e. Warburg effect) is one of the major metabolic alterations observed in cancer. This metabolic reprogramming is associated with the M2 isoform of Pyruvate Kinase (PKM2), a rate-limiting enzyme mainly expressed in embryonic and cancer cells. In normal tissues, the M1 isoform is expressed over M2 through alternative splicing of the mutually exclusive exons 9 and 10. Unlike the other isoforms, PKM2 exists in a quaternary-active and dimer-inactive glycolytic structure. The regulation of this oligomeric balance by protein-protein interaction pharmacologic modulation is a new therapeutic strategy for cancer treatment, particularly glioblastoma (GBM). The aim of this work is to discover novel selective compounds which would be able to modulate the oligomeric dynamics of PKM2. Selection of PKM2 structure was based on conformational diversity evaluation of all crystal structures using CoDNas database. Docking based Virtual Library Screening was performed using about 500.000 drug like compounds, focusing on key residues involved in the inter-subunit contacts and nuclear localization signaling. To enhance the likelihood of identifying a candidate, we conducted a chemical diversity analysis of the top ten ranked molecules using four different fingerprints, revealing a relatively high diversity among them. Subsequently, the biological activity of these compounds was assessed in GBM cell lines, indicating that 2 of them exhibited the ability to inhibit cell proliferation with an IC 50 ranging between 54.7-62.8 μM and 13.9-28.5 μM for compounds 3 and 8 respectively. To validate the molecular mechanism underlying this biological activity, we examined pyruvate kinase activity and PKM2 gene expression. Our results demonstrated a decrease in pyruvate kinase activity to 76.84% for compound 3 and 72.22% for compound 8 at the

lowest concentration tested (20μM). Additionally, treatment with compound 3 at 20 μM led to an approximately 80% decrease in PKM2 gene expression. Furthermore, a cross-linking assay revealed that treatment with these compounds reduced the abundance of the tetrameric form of PKM2. Experimental evidence suggests that two candidates identified through molecular docking have an effect on the growth of GBM cell lines. Therefore, the strategy evaluated in this study facilitated the identification of suitable candidates to modulate the oligomeric dynamics and specific activity of PKM2 in GBM. Further analysis of the candidate and its effects over other PKM2-related pathways will help us to determine its potential in the context of cancer diseases.

P34

imputomics 2.0: comprehensive R package for missing data imputation in proteomics data

Jarosław Chilimoniuk[1], Krystyna Grzesiak[1,2], Jakub Kołodziejczyk[1], Adam Krętowski[1], Michał Ciborowski[1], Michał Burdukiewicz[1,3]

[1]Clinical Research Centre, Medical University of Białystok, Białystok, Poland
[2]Faculty of Mathematics and Computer Science, University of Wrocław, Wrocław, Poland
[3]Institute of Biotechnology and Biomedicine, Autonomous University of Barcelona, Cerdanyola del Vallès, Spain

Background:
Missing data is a common issue in proteomics, which can negatively affect downstream statistical analyses and the interpretation of results. Missing values (MV) in proteomics can be divided into two types, missing not at random (MNAR) and missing at random (MAR). It is widely believed that the third MV type, missing completely at random (MCAR), is the same as MAR.
Various methods have been proposed to tackle this issue, from simple approaches like zero, mean, and median imputation to more complex techniques like random forest, singular value decomposition, and k-nearest neighbors. However, selecting the best method depends on the data type and research question.
Aim:
We aim to provide a comprehensive overview of the existing algorithms for imputing missing values in proteomics datasets, considering different patterns of missingness, such as MAR, MNAR, and mixtures.
Methods:
To address this issue, we systematically reviewed the literature on imputation methods for proteomics data. The selected imputation methods included basic techniques as well as advanced machine learning-based approaches.
Results:
We intend to expand imputomics, our R package and Shiny webserver originally developed for MV imputation in metabolomics datasets. We plan to incorporate identified and tested implementations of the MV imputation algorithms (MVIAs) that were used with proteomics datasets.

P35

UdonPred: Predicting protein disorder based on novel TriZOD scores

David Wagemann, Julius Schlensok, Tobias Senoner, Markus Haak, Burkhard Rost[1]

[1]Technical University of Munich, Germany

Intrinsically disordered proteins (IDPs) are a unique class of proteins that lack a fixed or rigid three-dimensional structure under physiological conditions. They are pivotal in various biological processes, including signal transduction, transcriptional regulation, and cell-cycle control. The study of IDPs is crucial as their disorder is often linked to diseases such as cancer, cardiovascular diseases, and neurodegenerative disorders.

While the advancement in computational methods has significantly enhanced the study of IDPs, many predictors available are limited by a shortage of high-quality disorder annotations. To improve upon existing methods, UdonPred uses a new dataset based on the TriZOD scoring scheme that assigns continuous disorder scores (G-scores) to residues based on nuclear magnetic resonance (NMR) spectroscopy chemical shifts in the entire Biological Magnetic Resonance Data Bank (BMRB), assessing how much experimentally determined chemical shifts deviate from random coil chemical shifts.

Based on this new dataset of over 15k peptides (10 times as large as the previously published CheZOD dataset), we trained slim neural networks to predict TriZOD G-scores from embeddings generated by the pre-trained protein language model (pLM) ProstT5. Needing only the information present in these embeddings, UdonPred itself is extremely fast and can predict per-residue disorder scores for a protein in a matter of milliseconds on a consumer-grade CPU.

UdonPred can successfully predict disorder from per-residue pLM embeddings: Evaluation on the commonly used OdinPred test set of 117 proteins exhibits promising results. With an area under the ROC curve (AUROC) of 0.907±0.01 between TriZOD G-scores predicted by UdonPred and manually computed binary CheZOD Z-scores, UdonPred is competitive with other state-of-the-art disorder predictors.

P36

Unrevealing residue conductors driving functional motion in proteins

Kemal Demirtaş, Türkan Haliloğlu[1]

[1]Chemical Engineering Department of Boğaziçi University

Allostery, characterized by an intricate network of allosteric interactions and their causal connections, emerges as a key biological regulatory mechanism. This study introduces a novel method that integrates the Gaussian Network Model (GNM) with Transfer Entropy (TE) calculations, utilizing solely sequence information. By dissecting dynamic information into distinct subsets of slow

modes, it reveals diverse layers of multi-directional allosteric pathways inherent within coevolutionary sequence data. Illustrated by exemplary cases, the scope and applicability of this approach from globular to non-globular proteins hint at its potential in identifying essential information source residues with the dynamic capacity to collectively influence others, thereby driving functional dynamics.

P37

Identification and Analysis of Cancer-Associated and Dysfunctional Non-Globular Proteins: A Bioinformatics Approach

Kevser Kübra Kırboğa

Bilecik Seyh Edebali University, Faculty of Engineering, Bioengineering Department, 11100, Bilecik, Türkiye

Aim and Background: Cancer continues to be a significant public health problem worldwide. Identification and analysis of cancer-associated proteins can help better understand the disease and identify potential therapeutic targets. This study focuses on the identification and analysis of cancer-related and dysfunctional non-globular proteins (NGP). NGPs play an important role in the development of many diseases such as cancer, due to their often disordered structure and broad functional properties. Detailed study of these proteins may help to better understand the biological processes associated with cancer. Therefore, this study provides a method for the identification and analysis of cancer-associated NGPs.
Method: This study aimed to identify and analyze cancer-associated and dysfunctional NGPs. First, NGPs were selected from the DisProt database. These proteins are biased to be cancer-associated and malfunctioning. Then, among these proteins, those larger than 300 amino acids were selected. This selection was made to study the functional properties and interactions of the proteins in more detail. Selected proteins were analyzed in STRING and PANTHER databases. These analyses have been used to understand the functional properties, interactions, and potential role of proteins. This method provides an effective approach to identifying cancer-associated and malfunctioning proteins and examining these proteins in more detail.
Results: The results of this study show that the identified NGPs are associated with various biological processes, molecular functions, cellular components and diseases. In terms of biological processes, processes such as tertiary branching, N-terminal peptidyl-lysine acetylation, glucocorticoid receptor signaling pathway, TORC2 signaling and regulation of cellular response to heat stand out in mammalian gland duct morphogenesis. In terms of molecular functions, pre-mRNA intronic binding, nuclear steroid receptor activity, transcript coactivator binding, nuclear receptor activity and damaged DNA binding are noteworthy. In terms of cellular components, F-actin coat protein complex, DNA repair complex, catalytic step 2 spliceosome, I banding, and cell-substrate junction are significant findings. In terms of diseases, proteins associated with childhood disintegrative disease, pelvic inflammatory disease, nominal aphasia, Rubinstein-Taybi syndrome and gynecomastia were identified. These findings constitute a valuable reference for further examination

of cancer-associated and dysfunctional NGPs and for evaluating these proteins as potential therapeutic targets.
Conclusion: This study provides a method for the identification and analysis of cancer-associated and dysfunctional NGPs. However, limitations of this study include that the databases and analysis methods used were focused on a specific set of proteins and therefore not all potential proteins were covered. Secondly, this study identified only proteins associated with specific biological processes, molecular functions, cellular components, and diseases. Therefore, other potential proteins and associated processes may have been overlooked.

P38

Unveiling Allosteric Modulation of 14-3-3 Proteins: Implications for Neurodegenerative Disease Research

Sergio Muller[1], Lautaro Rivera[1], Marina Uhart[1], Exequiel Barrera[1], Rostislav Skrabana[2], Diego Martin Bustos[1,3]

[1]Laboratorio de Integración de Señales Celulares, IHEM, Universidad Nacional de Cuyo, Mendoza, Argentina
[2]Institute of Neuroinmunology SAS, Bratislava, Slovakia
[3]FCEN, Universidad Nacional de Cuyo, Mendoza, Argentina

14-3-3 proteins, pivotal mediators in cellular signaling, orchestrate diverse biological processes through interactions with a plethora of phosphorylated partner proteins. Previous bioinformatic studies have suggested that besides phosphorylation, a common feature among recognized motifs on these targets is their high level of disorder, known as intrinsically disordered regions (IDRs). This characteristic enables the binding of multiple partners despite their structural differences. Additionally, these targets can adopt slightly different conformations as they interact with the binding site on the 14-3-3 protein. A recent breakthrough in our research has unveiled an allosteric site governing the conformational dynamics of 14-3-3 proteins, toggling between open and closed states. Through virtual screening of this region, we identified a family of compatible ligands named β-carbolines. As a first approach, we selected the parent β-carboline – Norharmane (NH). In a subsequent pull-down experiment, we investigated the influence of this compound on the binding of a well-known 14-3-3 partner, R18. Results indicated that R18 exhibited increased binding only when Norharmane was added post-incubation of 14-3-3 and R18. Conversely, pre-incubation of 14-3-3 with Norharmane obstructed R18's interaction. Given the association of 14-3-3 proteins with neurodegenerative diseases and their influence on the formation of neurofibrillary tangles, our current focus is on exploring the potential of β-carbolines for regulating these processes. Elucidating the influence of allosteric regulators on 14-3-3 activity could provide valuable insights into the intricate mechanisms underlying 14-3-3 protein regulation, particularly concerning non-globular proteins associated with neurodegenerative diseases.

P39

Mutations in the low-complexity domain of hnRNPA2 increase its phase separation and aggregation propensity

Duran-Armenta Luis Fernando[1,2], Lázár Tamás[1,2], Navarro Álvaro[3], Maes Dominique[2], Tompa Peter[1,2]

[1]VIB-VUB Center for Structural Biology, Vlaams Instituut voor Biotechnologie, Belgium
[2]Structural Biology Brussels, Vrije Universiteit Brussel, Belgium
[3]Structural Bioinformatics Unit, Fundación Instituto Leloir, Argentina

Heterogeneous ribonucleoprotein (hnRNP) A2 is a multifunctional RNA-binding protein that harbors an intrinsically disordered low-complexity domain (LCD). This region enables hnRNPA2 to undergo liquid-liquid phase separation (LLPS). Mutations in the LCDs of phase-separating proteins may disturb the physiological balance of LLPS, leading to the formation and accumulation of pathogenic aggregates. Notably, two mutations in the LCD of hnRNPA2 (D290V and P298L) have been associated with a range of (neuro)degenerative disorders encompassing amyotrophic lateral sclerosis (ALS), multisystem proteinopathy, and Paget's disease of the bone. However, the effects of these mutations on the phase separation behavior of hnRNPA2, as well as their contribution to aggregation, remain poorly understood. In this project, we assessed the evolutionary conservation of both D290 and P298 and the effect of mutations. Our bioinformatic analysis shows that the region encompassing D290 and P298 is highly conserved among vertebrates. This could explain why the mutations at these sites would lead to pathology, disturbing a well-established and functional sequence selected by evolution. In kinetic experiments, mutant hnRNPA2 LCD variants demonstrate a higher propensity for LLPS and aggregation, with P298L mutant displaying the highest tendency for aggregation, both in the presence and absence of prior LLPS. Both mutant variants form ThT-positive aggregates, indicating the presence of amyloid-like structures. Our goal is to further characterize the biophysical properties of the liquid droplets formed during LLPS and the resulting aggregates, to yield insights into the potential alterations induced by these mutations. We expect that these results will shed light on the underlying pathological mechanisms, contributing to a broader understanding of neurodegenerative disorders caused by aberrant phase transitions and protein aggregation.

P40

Significant combinations of amino acid pairs and triplets in low complexity regions in proteins

Joanna Ziemska-Legięcka[1], Aleksandra Gruca[2], Marcin Grynberg[1]

[1]Institute of Biochemistry and Biophysics PAS, Poland
[2]Silesian University of Technology

Low complexity regions (LCRs) are regions in proteins characterised by a high density of only a few types of amino acids. LCRs can be homorepeats, short tandem repeats, or simply strongly biased regions with fuzzy amino acid composition. For a long time, these regions were considered non-functional, but further research revealed its many functions. Even a manual inspection shows

that certain types of amino acids co-exist in sequences more often than others, so in this work, we decided to analyse this hypothesis.
To check the co-occurrence of amino acids, we performed statistical analyses on the dataset of LCRs from Uniref90. LCRs were identified with the SEG method using very restrictive parameters. This enabled us to find mostly repetitive regions, with low numbers of false positives. In the next step, we divided LCRs into datasets. Each dataset consisted of LCRs with one type of amino acid content greater than a given threshold. We used 10 thresholds for each type of amino acid in 10% intervals between values of 10% and 90%. Based on these datasets, we calculated hypergeometric tests on two or three datasets to check if these amino acids co-occurred more frequently together than alone. As a result, we showed that some amino acids tended to co-occur in LCRs, especially similar ones in their physico-chemical properties. This means that the same rule applied in reverse. We also presented examples of exclusion, when 2 amino acids stuck together only without the presence of a third amino acid, which co-occurred with one of the former in different conditions. Finally, we were able to explain such cases of exclusions with the co-occurrence of 3 amino acids in one LCR.

P41

Study of non-globular proteins present in the Occlussion Body of Spodoptera frugiperda granulovirus

Tomás Masson[1], Santiago Gomez Bergna[1], Matías L. Pidre[1], Ricardo Salvador[2], Víctor Romanowski[1], María Leticia Ferrelli[1]

[1]IBBM, UNLP-CONICET
[2]IMyZA, INTA

Spodoptera frugiperda granulovirus (SfGV) is capable of infecting and killing the fall armyworm (S. frugiperda), a significant insect pest in agriculture. This virus, along with other baculoviruses, is a very attractive biocontrol tools to be used as bioinsecticides for lepidopteran pests. They infect susceptible larvae by entering midgut epithelial cells after feeding on foliage contaminated with the virus. Baculoviruses exhibit two virion phenotypes depending on the infection stage: the Occludded Derived Virus (ODV) is responsible for the primary infection, while budded virions (BV) facilitate the dissemination of infection within the larval host. To effectively infect the host, ODVs must persist in favorable conditions in nature until a susceptible host encounters them while feeding. Subsequently, they must overcome the peritrophic membrane that protects midgut cells in larvae. To achieve this, baculoviruses occlude their ODV in Occlussion Bodies (OB) at the final stage of infection and release them when the insect dies and disintegrates. The OB is a cristalline structure formed by a major matrix protein (polyhedrin or granulin, depending on viral genus) and several other proteins. Given the crucial role of the OB in the viability and primary infection of baculoviruses, we aimed to investigate the protein composition of SfGV OB and assess the significance of each protein within the OB. We purified SfGV OB from S. frugiperda larvae and subjected them to mass spectrometry for protein detection. A total of seventy-two proteins encoded in SfGV genome were detected and assigned, based on previous knowledge of homologous proteins, either to the ODV structure, the OB matrix or with undefined locations. This work focuses on a set of non-globular proteins that are primarily short, unstructured, or possess an alpha helix as their main structural element. We explored

their function and relevance in the OB structure and/or infection. In particular, we discovered that ORF101, a highly abundant protein in the OB, with no detected pfam domains or homologous proteins of known function, was prone to aggregation when expressed in insect cells fused with GFP. Upon analyzing its sequence, we found out that it has a primary structure organization similar to P10, a well-studied baculovirus protein known to be involved in OB formation. This observation suggests a potential function associated with P10. This work highligths the importance of small non-globular proteins in the nature of baculoviruses and discusses their relevance in pest biocontrol and biotechnology.

P42

Evaluating the stability of polyXYs within IDRs with helical structure through Molecular Dynamics Simulations and AlphaFold Predictions

Mariane Goncalves-Kulik[1], Luis A. Baptista[2], Friederike Schmid[3], Miguel A. Andrade-Navarro[1]

[1]Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany
[2]Max Planck Institute for Polymer Research, Ackermannweg 10, 55128 Mainz, Germany
[3]Faculty of Physics, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany

Background. Intrinsically Disordered Regions (IDRs) are flexible regions of proteins that are known for presenting a reduced alphabet, most depleted of hydrophobic residues. IDRs often contain regions with a bias in their amino acid composition, known as low complexity regions (LCRs). While LCRs are common amongst IDRs little is known about their relationship with the formation or absence of secondary structures within these structures (Tompa, 2012; Uversky et al., 2000). The understanding of their folding mechanisms is important to explain their function (Uversky, 2010).
In our previous works, we evaluated the hypothesis that some of these LCRs could provide the IDR with a propensity to adopt a secondary structure. We studied LCRs composed of only one residue type (polyX) or two (polyXY). Targeting predicted IDRs in the human proteome, we initially focused on characterizing helical and extended folds through sequence homology with PDB structures (Gonçalves-Kulik et al., 2022) and later we expanded our search by evaluating the same secondary structures predicted by Alphafold (Gonçalves-Kulik et al., 2023). The findings for both works indicated that some polyXs (polyQ, polyE and polyK), and polyXYs (polyER, polyEK and polyED), show a propensity to adopt helical structures.
In this project, we propose to advance the understanding of why these specific amino acid compositions show a propensity to adopt helical structures. Applying Molecular Dynamics simulations (MD) and AlphaFold predictions without templates, we aim the exploration of the conformational landscape of these repeated regions, searching for the sequence patterns underlying these preferences.
Methods. The same steps executed in our previous work to annotate polyXY regions were performed (Gonçalves-Kulik et al., 2023), with updated IDR predictions and observed PDB evidences collected from MobiBD, along with the latest structure predictions from AlphaFoldV4. We then selected samples that present helical structures over the polyXY region in both AlphaFold predictions and PDB structures. MD simulations using CHARMM36IDPS force-field for 600ns were executed in a region of

100 residues surrounding the polyXY. Secondary structure propensities were then generated and each sample's characteristics were investigated.
Results and next steps. The structured regions of the 100 residues of TERA_HUMAN (Transitional endoplasmic reticulum ATPase) maintained a stable helical structure for the first 300ns of the simulation, however, between 300 and 600ns, the polyXY region show indications of unfolding. Our next steps include the extension of MD simulations, evaluation of regions with similar characteristics and selection of new candidates for simulations.

References
Tompa, P. Intrinsically disordered proteins : a 10-year recap. Trends in Biochemical Sciences 2012, 37, 509–516. https://doi.org/10.1016/j.tibs.2012.08.004.
Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why are 'natively unfolded' proteins unstructured under physiologic conditions? Proteins: Structure, Function and Genetics 2000, 41, 415–427. https://doi.org/10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7.
Uversky, V.N. The Mysterious Unfoldome: Structureless, Underappreciated, Yet Vital Part of Any Given Proteome. Journal of Biomedicine and Biotechnology 2010, 2010, 1–14. https://doi.org/10.1155/2010/568068.
Gonçalves-Kulik, M.; Mier, P.; Kastano, K.; Cortés, J.; Bernadó, P.; Schmid, F.; Andrade-Navarro, M.A. Low Complexity Induces Structure in Protein Regions Predicted as Intrinsically Disordered. Biomolecules 2022, 12. https://doi.org/10.3390/biom12081098.
Gonçalves-Kulik M, Schmid F, Andrade-Navarro MA. One Step Closer to the Understanding of the Relationship IDR-LCR-Structure. Genes. 2023; 14(9):1711. https://doi.org/10.3390/genes14091711

P43

Farming Protein Disorder Data with TriZOD by Leveraging NMR Spectroscopy

Markus Haak, Tobias Senoner, Burkhard Rost[1]

[1]Department of Informatics, Bioinformatics & Computational Biology - i12, Technical University of Munich (TUM), Garching/Munich, Germany

Accurate quantification of intrinsic disorder, crucial for understanding functional protein dynamics, remains challenging. We introduce TriZOD, an innovative scoring system for protein disorder analysis, utilizing nuclear magnetic resonance (NMR) spectroscopy chemical shifts. Traditional methods provide binary, residue-specific annotations, missing the complex spectrum of protein disorder. TriZOD extends the CheZOD scoring framework with quantitative statistical descriptors, offering a nuanced analysis of intrinsically disordered regions. It calculates per-residue scores from chemical shift data of polypeptides in the Biological Magnetic Resonance Data Bank (BMRB). The CheZOD Z-score is a quantitative metric for how much a set of experimentally determined chemical shifts deviate from random coil chemical shifts. The TriZOD G-scores extend upon them to be independent of the number of available chemical shifts. They are normalized to range between 0 and 1, which is beneficial for interpretation and use in training disorder predictors. Additionally, TriZOD introduces a refined, automated selection of BMRB datasets, including filters for physicochemical properties,

keywords, and chemical denaturants. We calculated G-scores for over 15,000 peptides in the BMRB, approximately 10-fold the size of previously published CheZOD datasets.
Validation against DisProt annotations demonstrates substantial agreement yet highlights discrepancies, suggesting the need to reevaluate some disorder annotations. TriZOD advances protein disorder prediction by leveraging the full potential of the BMRB database, refining our understanding of disorder, and challenging existing annotations.

P44

Unravelling the interactions of Nanobodies with pathologically relevant regions of Ataxin-3

Martyna Podlasiak, José Antonio Manso, Zsuzsa Sárkány, Pedro J B Pereira, Sandra Macedo Ribeiro[1,2]

[1]i3S–Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
[2]IBMC–Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal

Ataxin-3 (Atx-3) is an intrinsically disordered protein associated with the development of Machado-Joseph Disease (MJD). This protein comprises a well-defined, catalytic N-terminal domain, known as the Josephin Domain (JD), and a disordered C-terminal region. The C-terminal region encompasses three ubiquitin-interacting motifs (UIMs) alongside a critical polyglutamine tract (PolyQ) related to MJD pathogenicity.
The abnormal expansion of the PolyQ region initiates a protein aggregation mechanism that leads to cellular dysfunction and neurodegeneration, which are hallmarks of MJD. It was recently found that the interaction of Atx-3's UIMs with proteins (e.g., Hsc70-4) enhances Atx-3 aggregation and, thus, its pathogenicity.
In our lab, we have developed Atx-3-targeting nanobodies (NBs) that differently affect its aggregation. In this study, we employed an AI-driven predictive model and biophysical techniques, including isothermal titration calorimetry, to evaluate the interaction between NBs and Atx-3. Our primary objective was to gain insights into the mechanisms underlying their differential impact on aggregation. Interestingly, we found that the interaction between Atx-3 and NBs occurs primarily through Atx-3 UIM regions. Moreover, our observations indicate that there are various UIM-mediated binding modes among different NBs tested. These NBs interact with Atx-3 with affinities in the low-micromolar range and may offer a promising strategy to compete with UIM-mediated pathological Atx-3 interactions. Overall, our results provide valuable insights into the mechanisms that govern Atx-3 aggregation and offer new tools for novel disease-modifying therapies.

P45

Decoding Autophagy: An Integrative Approach to study Protein-Protein Interactions

Mattia Utichi[1,2] , Matteo Arnaudi[1,2], Oana Nicoleta Antonescu[1], Luca DI Leo[3], Valentina Sora[1,2], Mukesh Kumar[1], Matteo Tiberti[1], Emiliano Maiani[1], Matteo Lambrughi[1], Daniela De Zio[3], Elena Papaleo[1,2]

[1]Cancer Structural Biology, Danish Cancer Institute, Copenhagen, Denmark

[2]Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark, Lyngby, Denmark
[3]Melanoma Research Team, Danish Cancer Institute, Copenhagen, Denmark

Autophagy, a conserved cellular process, serves as the cell's recycling system. It coordinates the degradation of damaged organelles, protein aggregates, and macromolecules, ensuring cellular balance and energy supply. Dysregulation of autophagy contributes to various diseases, including cancer, neurodegenerative disorders, and infections. Protein-protein interactions play a central role in these pathological processes. Studying these interactions provides insights into disease mechanisms and potential interventions.

Notably, autophagy heavily relies on Intrinsically Disordered Proteins (IDPs). Autophagy proteins often harbor LC3-interacting motifs (LIRs), enabling interactions with essential autophagy components. These motifs facilitate selective cargo recognition and autophagosome formation. Investigating IDPs poses unique hurdles due to their inherent flexibility, and their dynamic behavior necessitates a multifaceted approach. In this study, we present our integrative approach, combining unbiased molecular simulations, biophysical assays, and cellular experiments, to study the complex network of autophagy interactions and their regulation.

We specifically investigated the interactions of MAP1LC3B, a key component of the autophagy machinery, with three different IDPs: FYCO1, OPTN, and AMBRA1.

FYVE And Coiled-Coil Domain Autophagy Adaptor 1 (FYCO1) acts as an autophagic adaptor protein involved in the microtubule transport of autophagosomes. It facilitates the movement of these vesicles within the cell. Through a combination of molecular dynamics simulations, computational mutational scans, and cellular experiments, we uncovered the importance of the flanking regions of FYCO1 LIR and the possible post-translational regulation of its interaction with MAP1LC3B through phosphorylation of T1248.

Optineurin (OPTN) acts as a well-recognized autophagy receptor. It participates in cargo recognition, autophagosome formation, autophagosome maturation, and lysosomal quality control. Our investigation focused on the impact of multiple phosphorylation events in close proximity to its LC3-interacting region (LIR). We employed enhanced sampling molecular simulations and biophysical assays, including Microfluidic Diffusional Sizing (MDS) and Surface Plasmon Resonance (SPR).

Activating molecule in Beclin1-regulated autophagy (AMBRA1) plays a positive role in the formation of autophagosomes during mitophagy, a specific type of selective autophagy targeting mitochondria. Through a combination of sequence analysis, molecular simulations, and cellular approaches, we discovered the existence of multiple LIRs in AMBRA1.

The insights gained from these case studies have paved the way for the development of an experimental and computational mixed workflow for investigating protein-protein interactions and protein function in LIR-containing proteins. This approach holds value not only in the specific context of autophagy but could also be extended to other proteins containing Short Linear Motifs (SLiMs).

P46

Confident dataset generation for the specific prediction and benchmarking of client, driver and negative proteins in liquid-liquid phase separation

Michał Burdukiewicz[1,2], Carlos Pintado-Grima[2], Oriol Bárcenas[2], Valentín Iglesias[2], Salvador Ventura[2]

[1]Medical University of Białystok, Poland
[2]Autonomous University of Barcelona, Spain

Proteins self-organize in dynamic cellular contexts by assembling into reversible biomolecular condensates by liquid-liquid phase separation (LLPS). Such condensates can be constituted by one or many proteins, which may have different roles in the final supramolecular structure. While driver proteins form condensates by themselves without the requirement of any partner, client proteins are later recruited into them and are not essential for their integrity. Although several databases already collect proteins undergoing LLPS, each resource contains different data types. Moreover, the lack of proteins without any experimental association with condensates (negative data) prevents the generation of confident predictors and fair benchmarks. To alleviate this issue, we have inspected all relevant LLPS databases and applied rational filters to generate datasets of client and driver proteins, where multiple information sources support the annotations with a similar confidence level. Besides, we generate negative datasets, including both globular and disordered proteins. Considering that disorder can overlap with condensation in some instances, we have further split all datasets into shorter disordered-related elements to avoid sequential biases. Interestingly, in a first analysis with such datasets, we observed that, for specific physicochemical properties related to LLPs, certain disordered elements, such as the presence of cryptic amyloidogenic regions (CARs), better identify significant differences when compared with negative data. Our datasets could be used to train models for attaining better predictive performances, build more standardized benchmarks, and identify possible physicochemical differences between LLPS-positive and LLPS-negative proteins.

P47

Phase-separating fusion proteins drive cancer by dysregulating transcription through ectopic condensates

Nazanin Farahi, Tamas Lazar, Peter Tompa, Bálint Mészáros, Rita Pancsa*

Numerous cellular processes rely on biomolecular condensates formed through liquid-liquid phase separation (LLPS), thus, perturbations of LLPS underlie various diseases. We found that proteins initiating LLPS – scaffolds – are frequently implicated in somatic cancers, even surpassing their involvement in neurodegeneration. Cancer-associated LLPS scaffolds are connected to all cancer hallmarks and tend to be oncogenes with dominant genetic effects currently lacking therapeutic options. Since most cancer-associated LLPS scaffolds exert their oncogenic effects as oncogenic fusion proteins (OFPs), we undertook a systematic analysis of cancer driver OFPs by assessing their module-level molecular functions. We identified both known and novel combinations of molecular

functions that are specific to OFPs and thus have a high potential for driving tumorigenesis. Protein regions driving condensate formation show an increased association with DNA- or chromatin-binding domains of transcription regulators within OFPs, indicating a common molecular mechanism underlying several soft tissue sarcomas and hematologic malignancies where phase-separation-prone OFPs form abnormal, ectopic condensates along the DNA, and thereby dysregulate gene expression programs.

P48

Insights into the Alzheimer's disease specific pre-aggregation conformation of monomeric tau proteins using antibodies DC11 and MN423

Ondrej Cehlar[1], Stefana Njemoga[1], Rostislav Skrabana[1], Viliam Volko[2], Jozef Hritz[2,3], Pavel Kaderavek[2], Branislav Kovacech[4]

[1]Institute of Neuroimmunology, Laboratory of Structural Biology of Neurodegeneration, Slovak Academy of Sciences, Bratislava, Slovakia
[2]CEITEC MU, Brno, Czech Republic
[3]Faculty of Science, Department of Chemistry, Masaryk University, Brno, Czech Republic
[4]Institute of Neuroimmunology, Structural Proteomics, Slovak Academy of Sciences, Bratislava, Slovakia

A key yet unresolved question of the pathogenesis of Alzheimer's disease (AD) and other tauopathies is the cause and the mechanism of the transition from the unstructured monomeric tau protein to the insoluble filaments deposited in the brain tissue. In the physiological state, tau protein exists as a conformational ensemble of interconverting structures and on the scale of transition from monomeric through oligomeric and filamentous species we can observe conformations reacting with specific antibodies, mainly with DC11, which is able to specifically discriminate between tau proteins isolated from healthy brain and tau proteins isolated from the brain of AD patient. The antibody recognizes also the recombinant truncated tau proteins up to the shortest fragment tau321-391 [1].
It was found that conformational antibodies DC11 and MN423 have catalytic pro-aggregatory effects in tau aggregation assay, whereas the antibody DC8E8 has inhibitory effects on tau filament formation [2]. This may imply possible mechanism of induction of pathological tau conformation, in which the antibody prepared against pathological tau imprints the pathological conformation into the physiological tau proteins in solution and therefore speeds up the tau aggregation.
To further uncover the binding mode of the conformational antibody DC11, we have performed NMR epitope mapping using 13C, 15N labelled tau321-391 and tau297-391 (dGAE). The overlay of HSQC spectra showed the region of tau between residues 370-390 to be affected by the binding of DC11, i.e. its C-terminal region. However, previous studies suggest the importance of region 321-325 for the interaction of tau with DC11 antibody. We have further characterized the influence of DC11 Fab binding on the non-epitope tau residues using NMR relaxation measurements of 15N labelled tau dGAE.

The results highlight the importance of the R' region of tau, that was recently shown to be important also for tau interaction with microtubules [3]. This sequence forms the interface of rigid filament core and flanking fuzzy C terminal segment in solved tauopathy filaments.

1. Vechterova, L.; Kontsekova, E.; Zilka, N.; Ferencik, M.; Ravid, R.; Novak, M. Neuroreport, 14 (2003), 87-91.
2. Kontsekova, E.; Zilka, N.; Kovacech, B.; Skrabana, R.; Novak, M. Alzheimers Res Ther, 6 (2014), 45.
3. El Mammeri, N., Dregni, A. J., Duan, P., Wang, H. K., Hong, M. Sci. Adv., 8 (2022), 4459.

Acknowledgment. This work was supported by research grants APVV 21-0479, VEGA 2/0125/23 and 2/0141/23, and by iNEXT-Discovery, grant number 871037, and MSCA-RISE 873127, funded by Horizon 2020 program of EC.

P49

Advances in High-Purity Extracellular Vesicle Protein Isolation for Neurological Biomarker Detection

Oscar Bello[1,2], Diego Bustos[2], Rostislav Skrabana[1], Estela Muñoz[2]

[1]Institute of Neuroimmunology, Slovak Academy of Sciences, Slovakia
[2]Instituto de Histología y Embriología de Mendoza, Universidad Nacional de Cuyo, Mendoza, Argentina

Extracellular vesicles (EVs) are known to carry protein components reflective of their cellular origins. Their capability to cross the blood-brain barrier positions them at the forefront of intense research, particularly due to their potential to provide insights into the biochemical status of various organs, the brain included. Characterizing the protein content of EVs may allow the identification of organ-specific biomarkers, which has significant implications for the early detection and monitoring of neurodegenerative diseases, such as tauopathies and synucleinopathies. The progress in this field has been hampered by the challenge of isolating EVs with high purity. Conventionally, over 99% of proteins within EVs fractions are serum proteins, inadvertently co-purified during the isolation process, which can obscure the particular protein markers sought after. This poster introduces an innovative method designed to overcome this challenge. Our technique facilitates the extraction of EVs with markedly higher purity and yield from blood samples. Importantly, the method is engineered to be implementable in diagnostic laboratories with basic equipment, thereby broadening access to this sophisticated diagnostic tool. It enables routine, non-invasive monitoring of brain health and could facilitate earlier intervention for neurological disorders.

P50

The Impact of Intrinsically Disordered Effector Regions on the developmental functions and DNA Binding Specificity of Drosophila Transcription Factor Bicoid

Dilan Akdogan, Pinar Onal[1]

[1]Bilkent University, Turkey

Precise spatio temporal regulation of gene expression is critical for robust developmental processes. Divergence in the regulation results in phenotypic variations seen between individuals as well as between diseased and healthy tissues. Transcription factors regulate transcription spatially and temporally by binding to 6-12 bp-long specific DNA sequences called motifs found usually upstream of the genes to be regulated. A specific eukaryotic transcription factor binds only a subset of the motifs which are abundant across the genome. How a transcription factor finds and binds to a specific motif at a specific time and space is not completely understood.

Drosophila maternal transcription factor Bicoid (Bcd) is an essential factor for embryo development. Through the anterior to posterior gradient it enables axis determination. It also organizes anterior structures such as head and thorax via spatio-temporal regulation of hundreds of genes involved in segmentation. Bcd uses its highly conserved DNA-binding Homeodomain, which has a Helix-Turn-Helix structure factor, and its effector domains that are relatively less conserved and that don't have a stable structure. Previously, we and others showed the significance of the Bicoid's activator domains for its developmental functions. However, the mechanism of action of these different effector domains is still to be addressed.

To investigate the details of mechanisms of action of these different effector regions in development and differentiation we use Drosophila melanogaster transcription factor Bcd. With this project, we propose first to systematically delete its effector domains of different features and generate transgenic flies whose sole source of Bcd is these transgenic variants and examine their contribution to embryo survival, segmentation and expression of Bcd target genes along the AP axis. Secondly, we aim to investigate the contribution of these different regions to DNA binding diversity and specificity of Bcd in live embryos. We expect to see different effector regions to contribute differentially to Bcd's developmental functions and to its DNA binding activity.

P51

A machine learning model for discovering repurposed drug candidates to inhibit the non-globular CTNNB1 protein within the Wnt/beta-catenin pathway in cancer

Selahattin Aksoy, Talip Zengin, Tugba Onal-Suzek

CTNNB1, also known as beta-catenin, displays a non-globular property attributed to its intrinsically disordered regions. These regions lack a defined three-dimensional structure, imparting flexibility and versatility to beta-catenin's functions. Recent research has delved into comprehending the implications of this non-globular nature in drug targeting strategies. Dysregulated Wnt signaling, propelled by mutations in CTNNB1, is implicated in tumor initiation and the maintenance of a cancer stem cell-like phenotype. It can promote the survival and growth of cancer cells, as well as resistance to certain treatments. To comprehend the role of CTNNB1 in cancer Wnt signaling, we

implemented a machine-learning pipeline to repurpose new candidate compounds for the non-globular CTNNB1 protein.

For this purpose, we randomly selected PubChem compounds through under-sampling from all the compounds active against CTNNB1 in the PubChem BioAssay database. A total of 144 compounds were used for training, evenly split between 72 positive and 72 negative classes. Ten-fold cross-validation was applied to the training set to select the optimal model for parameter tuning. Following cross-validation, the top 5 models selected for parameter tuning were: k-nearest neighbor, Ada-Boost classifier, Random Forest, Extra Trees, and Naive Bayes. Their respective accuracy scores were 0.83, 0.83, 0.82, 0.82, and 0.82. Following parameter tuning and subsequent testing on a set of 16 compounds, comprising 8 positive and 8 negative compounds, the Random Forest model was selected as the top performer. The model achieved an accuracy of 0.83, an AUC of 0.87, and an F1 score of 0.83 when evaluated on the test set of compounds. Finally, the Random Forest model, trained on the complete dataset consisting of 144 training compounds and 16 test compounds (totaling 160 instances with 80 positive and 80 negative samples), was evaluated on a holdout dataset containing 9 positive and 9 negative compounds. The model achieved an accuracy of 0.83, an AUC of 0.83, and an F1 score of 0.82 over this previously unseen dataset.

Ultimately, we utilized the model to conduct a virtual screening of the entire PubChem Compound database, encompassing 110 million compounds downloaded in 2022. All the predicted compounds are sorted by their prediction scores. One of the top 100 compounds, CID126740043, is validated by its 90% fingerprint similarity to CID122632882, which has patented evidence of CTNNB1 inhibition.

P52

A3D Model Organism Database (A3D-MODB): a database for proteome aggregation predictions in model organisms

A.E. Badaczewska-Dawid[1], A. Kuriata[2], C. Pintado-Grima[3], J. Garcia-Pardo[3], M. Burdukiewicz[3,4], S. Ventura[3], S. Kmiecik[2], V. Iglesias[3,4]

[1]Genome Informatics Facility, Office of Biotechnology, Iowa State University, Ames, 50011 IA, USA
[2]Biological and Chemical Research Center, Faculty of Chemistry, University of Warsaw, Pasteura 1, 02-093 Warsaw, Poland
[3]Institut de Biotecnologia i de Biomedicina (IBB) and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain
[4]Clinical Research Centre, Medical University of Białystok, Kilińskiego 1, 15-369, Białystok, Poland

Protein aggregation has been associated with aging and different pathologies and represents a bottleneck in the industrial production of biotherapeutics. Numerous past studies performed in Escherichia coli and other model organisms have allowed to dissect the biophysical principles underlying this process. This knowledge fuelled the development of computational tools, such as Aggrescan 3D (A3D) to forecast and re-design protein aggregation. Here, we present the A3D Model Organism Database (A3D-MODB) http://biocomp.chem.uw.edu.pl/A3D2/databases, a

comprehensive resource for the study of structural protein aggregation in the proteomes of 12 key model species spanning distant biological clades. In addition to A3D predictions, this resource incorporates information useful for contextualizing protein aggregation, including membrane protein topology and structural model confidence, as an indirect reporter of protein disorder. A3D-MODB is openly accessible without need for registration. We anticipate the A3D Model Organism database becoming an important platform for conducting in-depth multi-species aggregation analyses.

P53

Putative intrinsically disordered DNA-binding proteins in algal pyrenoids

Viktoriia Komarysta

Department of Botany and Plant Ecology, School of Biology, V.N. Karazin Kharkiv National University, Kharkiv, Ukraine

Pyrenoid is an ancient sub-organelle in chloroplasts of many algae and some hornworts, a phase-separated liquid proteinaceous body inside the chloroplast stroma, lacking in the other evolutionary advanced plants. Pyrenoid predominantly consists of RuBisCo, the key enzyme of photosynthetic CO2 assimilation, carbonic anhydrase, and several hundreds of other proteins. Highly intrinsically disordered protein EPYC1 binds RuBisCo and retains it inside the pyrenoid. Compartmentalization of RuBisCo, its isolation from O2-evolving photosystem II, and local CO2 concentrating due to carbonic anhydrase activity mitigate photorespiration and enhance photosynthetic effectiveness in pyrenoid-containing algae. That is considered the principal function of pyrenoid.
There is a significant amount of microscopy data showing the localization of chloroplast DNA in pyrenoids of algal chloroplasts, at least in some cell cycle phases, among them our observation in Dunaliella salina during chloroplast division. Neither the mechanism nor biological role of this phenomenon is understood.
The hypothesis behind this study is that chloroplast DNA specifically binds to some proteins in the algal pyrenoid. These proteins could be highly intrinsically disordered, thus holding chloroplast DNA inside the pyrenoid.
To test the hypothesis, we inferred the Chlamydomonas reinhardtii pyrenoid proteome from mass spectrum data on isolated pyrenoids available in the PRIDE database (PXD004509) using SearchGUI and PeptideShaker protocol on the European Galaxy Biostars server. The Eukaryotic Nucleic Acid Binding Protein Database (ENPD) was used to cross-check the list of pyrenoid proteins and identify those that bind to DNA. The matched proteins were subjected to the IUPred2A tool to identify highly intrinsically disordered molecules among DNA-binding proteins of algal pyrenoids.
DNA-binding proteins presented a small fraction of the total pyrenoid proteome (23 proteins). About half of them, nuclear histones and histone variants (11), could result from contamination of pyrenoid fraction due to the methodological problems of pyrenoid isolation. Some revealed proteins were subunits of chloroplast DNA-directed RNA polymerase (5), reparation factors (2: DNA-ligase, recombination protein RecA), and transcription regulators (3: Whirly, mTERF, ARID domain-containing protein), with highly ordered molecular structures. The remaining were high mobility group box (HMG-box) domain-containing proteins (2: A8HXE1, A8J3F0) with a relatively high level of intrinsic

molecular disorder, especially in A8HXE1, making this protein a candidate to promote chloroplast DNA localization into pyrenoid.
Research on the interaction between DNA and pyrenoid in algae may have implications for understanding chloroplast evolution. Intrinsically disordered pyrenoid proteins that bind DNA could serve as an experimental model for studying this class of proteins.

P54

Unraveling the Role of Intrinsically Disordered Proteins in Reproductive Biology

Ximena Aixa Castro Naser[1], Zarifa Osmanli[2], Martin E Argañaraz[1], Silvio C E Tosatto[2], Alexander Miguel Monzon[2]

[1]Instituto Superior de Investigaciones Biológicas, San Miguel de Tucumán, Argentina
[2]Department of Biomedical Sciences, University of Padova, Padova, Italy

Embryo implantation in the uterine endometrium represents a crucial step in mammalian reproduction, requiring intricate signaling interactions between maternal endometrium and fetal trophoblast cells. The involvement of intrinsically disordered proteins (IDPs) adds complexity to these pathways due to their unique structural properties. IDPs lack a well-defined three-dimensional structure under physiological conditions, enabling them to engage in multiple protein interactions crucial for various cellular processes, including signaling, transcriptional regulation, and molecular recognition. Dysregulation of IDPs can lead to a spectrum of human diseases, including cancer, cardiovascular diseases, and infertility.
In this study, our aim was to elucidate the role of IDPs in reproductive events. Drawing from mass spectrometry data of endometrium and pre-implantation embryos, we employed several bioinformatic tools, such as MobiDB and CAID, to identify and characterize the intrinsic disorder of proteins in these samples, while also conducting in-depth analyses of their sequence, structure, and function. Functional annotation tools, including Gene Ontology, and enrichment analysis were utilized to elucidate the involvement of IDPs in crucial reproductive processes such as maternal-embryonic dialogue and implantation.
Furthermore, we explored the potential interactions between IDPs and other proteins involved in reproduction using resources like the ELM and STRING databases. Network analysis was performed using Cytoscape to identify key interacting partners associated with IDPs. Additionally, we investigated the evolutionary conservation and diversity of IDPs across different mammalian species using MEGA software to identify conserved regions or motifs and to understand the evolutionary history and functional divergence of IDPs in reproductive processes.
Our findings shed light on the functional significance of IDPs in maternal-embryonic dialogue, embryonic development, and implantation, thereby deepening our understanding of reproductive biology. They also inspire new lines of research in diagnostic and therapeutic interventions in reproductive medicine.

P55

An Integrative structure-based approach to link diseases associated mutations to pathogenicity

Matteo Arnaudi[1,2], Emiliano Maiani[2], Karolina Krzesinska[1], Giacomo Miletti[2], Birthe B Kragelund[3], Mattia Utichi[1,2], Elena Papaleo[1,2], Apolinar Mendoza[2]

[1]Danish Technical University, Denmark
[2]Danish Cancer Institute, Denmark
[3]University of Copenhagen, Denmark

Pathogenicity annotations of germline mutations obtained through NGS techniques represent a significant advancement for healthcare and personalized medicine, in particular for disease treatment, genetic counseling, prevention, and monitoring. However, most part of them are Variants of Unknown Significance (VUS), which hinder patient care and therapeutic development due to challenges in determining their pathogenicity. Advances in Artificial Intelligence have led to the development of highly accurate pathogenicity classifiers for VUS classification. Nevertheless, these classifiers provide only probability scores, lacking the molecular mechanisms underlying mutation pathogenicity and limiting their clinical utility. This emphasizes the need for comprehensive interpretations essential for clinicians and molecular biologists.
To address this gap, we are developing a structure-based framework that integrates computational and experimental methodologies to predict and validate the effect of missense germline mutations associated with various diseases on protein structural features. We retrieve target protein structures through AlphaFold and/or the Protein-DataBank (PDB) database. We trim predicted structures using disorder propensity information obtained through a combined approach involving AF pLDDT score and CAID. We employ experimental databases of manually curated complexes for the selection of potential interactors and protein sequence analysis for the Short Linear Motifs (SLiMs) annotations. We combine the information to generate the complexes through AF-multimer in the absence of experimental structures.
We obtain structural ensembles through Molecular Dynamics simulations, to include protein-flexibility and improve prediction accuracy. We integrate free and binding energy calculations, sequence analysis, and Structure-Based-Statistical-Mechanical-Models to predict the effects of mutation collected from three different databases on protein stability, protein-protein interactions, changes in post-translational modification sites, and allosteric changes on protein pockets. These predictions guide experimental validation through a platform currently under development that integrates biochemical and microscopy techniques.
In our preliminary experiments we validated the destabilizing effect of the P263H variant on ARID3A, a transcriptional factor highly mutated in leukemia, and its interaction with the DNA damage sensor TP3BP1 through CHX assay and microscopy techniques. Additionally, we validated sequence analysis predictions identifying a SH3-mediated interaction between Lyn tyrosine kinase and P14arf, suggesting further research avenues in B-cell cellular pathways. These results, along with the computational workflow, not only provide a robust method for investigating mutation effects on protein structural features but also demonstrate potential for exploring intrinsically disordered proteins (IDPs), particularly SLiM interactions.
Overall, this framework supports molecular biologists in both basic and translational research, aiding in the understanding of disease mechanisms and the development of new drugs and treatments.

P56

Structural Characterization of Amyloid Fibrils of Tau Variants Using Atomic Force Microscopy

Zuzana Bednarikova[1], Ondrej Cehlar[2], Stefana Njemoga[2,3], Zuzana Gazova[1]

[1]Institute of Experimental Physics Slovak Academy of Sciences, Kosice, Slovakia
[2]Institute of Neuroimmunology Slovak Academy of Sciences, Bratislava, Slovakia
[3]Faculty of Natural Sciences, Comenius University, Bratislava, Slovakia

The formation and accumulation of amyloid fibrils represent key pathological features in over 50 human disorders, including neurodegenerative diseases and systemic amyloidoses. Despite significant research, much remains unknown regarding the relationship between the structural elements of fibrils and the pathology of amyloid-related diseases. While techniques such as cryo-EM and solid-state NMR offer high-resolution structural characterization of amyloid fibrils, they generally provide structural information of the average fibril ensembles in the sample under study. In this study, we present a structural characterization of individual populations of Tau variant amyloid fibrils using atomic force microscopy (AFM). The amyloid fibrils were formed in vitro from five fragments of tau (297–391, 306-391, 316-391, 321-391, 326-391). The fibrils formation was monitored using Thioflavin T assay, and resulting fibrils were visualized using NTegra atomic force microscope equipped with SNL-10 cantilever in air. Our data reveal that truncation of the tau sequence leads to the formation of fibrils with distinct morphologies and characteristics, with one tau variant failing to form amyloid fibrils altogether. Interestingly, one tau variant formed ThT-negative amyloid fibrils. These findings underscore the utility of individual particle structural analysis using AFM in providing valuable insights into the polymorphism of amyloid fibrils. Overall, our study sheds light on the diverse structural features of Tau variants amyloid fibrils.
This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-22-0598 and APVV-21-0479; Slovak Grant Agency VEGA 02/0176/21 and VEGA 02/0125/23; and COST Action ML4NGP and MVTS grant ML4NGP Slovakia.

Co-organizing Consortia

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823886.

Funded by the European Union

This event is part of the activities of the COST Action ML4NGP, CA21160, supported by COST (European Cooperation in Science and Technology).