Analysis of conformation manifolds for intrinsically disordered proteins

PhD advisor: Thérèse Malliavin (Laboratoire de Physique et Chimie Théoriques, Université de Lorraine), therese.malliavin@univ-lorraine.fr

PhD international co-advisor: Jung-Hsin Lin (Academia Sinica), jhlin@gate.sinica.edu.tw

PhD co-advisor: Jérôme Idier (Laboratoire des Sciences du Numeriques de Nantes), jerome.idier@ls2n.fr

Location: Laboratoire de Physique et Chimie Théoriques, Université de Lorraine

Searched profile

Master-completed student or equivalent, with a good background in data science. Good programming skills are required with a strong interest for applications to biological systems.

The PhD student is expected to make two stays of 6 months to Taiwan for developing the collaboration in the frame of the TIGP-X Pilot program, initiated in 2023 by Academia Sinica, and which will provide financial support for the travel and the stay.

Scientific Description

The PhD project intends to develop numerical approaches for improving the analysis of the conformational space of intrinsically disordered proteins. These approaches will be developed in the frame of a discrete method for enumerating protein conformations, the interval Branch-and-Prune (iBP) approach, developed by the PhD advisor at the University of Lorraine. Methods coming from the optimal transport theory, based on the use of Wasserstein distance, will be applied to design a clustering approach for describing the conformational space of intrinsically disordered proteins. The developed numerical approaches will be tested on two flexible and disordered proteins experimentally studied at Academia Sinica, Taiwan.

The enormous development of structural biology has permitted to determine more than 213,000 biomolecular structures at atomic resolution (Protein Data Bank: www.rcsb.org). These structures, defined by the relative positions of atoms, have been permitting strong advances in understanding the function of the protein molecules, and consequently in understanding essential physiological and pathological processes. The knowledge of protein structures is also instrumental for learning how to interfere with such processes, in particular by designing more powerful antibodies, or small compound inhibitors. Determining the structure of proteins is therefore a key aspect to progress in better management of health and biotechnology problems.

Along the last four decades, the development of biochemistry and structural biology has revealed that the proteins are much more mobile than what was expected from the first structures obtained using X-ray diffraction in the fifties. In addition, the discovery of amyloid peptides pointed out the importance of disorder in pathological processes and about twenty years ago, the importance of intrinsic disordered proteins in normal physiological processes was recognized. Nowadays, it is estimated that disordered residues constitute 35 to 50% of the human proteome and, depending on the organism type, the overall percentage of amino acids predicted to be disordered ranges from about 12% up to 50% [12].

The classical methods of calculations for protein structure are based on optimization techniques. While optimization is a rich research field currently providing several methods and algorithms for effective searching of minima, the roughness of the objective functions arising in our context makes it impossible to have guarantees of global optimality. In addition, in the case of IDRs or IDPs, however, the number of possible conformations is much bigger and they are in permanent interconversion, due to the flat free energy surface of the system. In this case, methods that allow for systematic enumeration are potentially capable of providing a path to a better description of the system, while avoiding bypassing any conformational region. The interval Branch-and-Prune (iBP) approach [9] allows a systematic enumeration of conformations in the frame of the distance geometry problem. This approach was adapted to the protein molecular modeling as threading-augmented interval Branch-and-Prune (TAiBP: [6, 11]), providing a framework for the systematic enumeration of protein conformations. TAiBP was shown to make possible the analysis of the conformational space of Intrinsically Disordered Region [10] or of Intrinsically Disordered Protein [7].

After generation of the protein conformations using TAiBP, these conformations the protein conformations are filtered using Small Angle X-ray Scattering (SAXS) curves or Ramachandran probability maps in order to determine representative conformations of the system along with their populations. Bayesian [8] or Gaussian mixture [7] approaches have been proposed for such purpose. The Gaussian mixture approach, developped by Jérôme Idier and Thérèse Malliavin [7] determines representative conformations and their corresponding populations using the Kullback–Leibler divergence in order to compare the probability distributions for the backbone torsion angles ϕ and ψ. Nevertheless, this divergence is appropriate if the probability laws have

the same support. The purpose of the PhD is to use the Wasserstein distance, derived from the theory of optimal transport, for modeling the conformational space of proteins. Such ideas have been successfully used recently for clustering conformational ensembles of intrinsically disordered proteins [4]. Nevertheless, this approach is used on full chain of disordered proteins. The present PhD project thus intends to adapt the use of the Wasserstein distance to the hierarchical procedure employed for generating the conformations during the TAiBP procedure.

The methods developed in the project will be first applied on intrinsically disordered proteins deposited in the PED database (proteinensemble.org) [3]. Then, the project will focus on two experimental examples of application, being two intrinsically disordered proteins which are studied since one year or will be studied by Jung-Hsin Lin at Academia Sinica. First, the small EDRK-rich factor 1 (SERF1a) is a 110-residue protein, involved in protein aggregation and is supposed to play an important role in the diseases such as Alzheimer's, Parkinson's, and Huntington's. It was shown [2] that the isolated SERF1a is predominantly disordered and that the interaction of SERF1a with the α-synuclein facilitates the conversion of α-synuclein monomers into amyloid fibers. A more recent study [5] points out the charge complementation between the different partners as key factors for establishing interaction.

The developed procedure will be also tested on the promyelocytic leukemia protein (PML), a major component of PML nuclear bodies (PML-Nbs), involved in key regulation activities, such as DNA replication, transcriptional regulation, cell cycle control, antiviral defense, apoptosis and tumor suppression. PML-NBs serves as a scaffold harboring numerous proteins and regulating their functions in response to cellular biogenesis. The function of PML-NBs is significantly associated with the SUMOylation process, where SUMO stands for Small Ubiquitin related MOdifier. PML has three major SUMOylation sites and one SUMO-interacting motifs (SIMs). PML SUMOylated sites are located at defined positions, known as the TRIpartite motifs (TRIM). However, the PML-SIM regions are intrinsically disordered and bind with SUMO through non-covalent interactions. Upon interacting with SUMO, PML-SIMs undergo transition towards β-sheets, as observed from recent X-ray crystallographic study [1]. We plan to harness the NMR and SWAXS data, along with the iBP conformation exhaustive enumeration and extensive molecular dynamics simulations, to understand the structural variations of peptides containing PML-SIMs in the presence and absence of SUMO protein. Furthermore, we would like to elucidate the structural bases of selectivity for interaction of PML with SUMO paralogs (SUMO1, SUMO2 and SUM03). Indeed, SUMO2 and SUMO3 share 97% sequence identity, whereas SUMO1 is 47% identical to SUMO2/3, and PML-SIMs interacts with SUMO1, rather than with SUMO2/3.

References

[1] Cappadocia et al. Structural and functional characterization of the phosphorylation-dependent interaction between PML and SUMO1. *Structure*, 23:126–138, 2015.

[2] Falsone et al. SERF protein is a direct modifier of amyloid fiber assembly. *Cell Rep*, 2:358–371, 2012.

[3] Ghafouri et al. PED in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic Acids Res*, page in press, 2023.

[4] González-Delgado et al. WASCO: A Wasserstein-based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins. *J Mol Biol*, 435(14):168053, 2023.

[5] Pras et al. The cellular modifier MOAG-4/SERF drives amyloid formation through charge complementation. *EMBO J*, 40:e107568, 2021.

[6] Worley et al. Tuning interval Branch-and-Prune for protein structure determination. *Journal of Global Optimization*, 72:109–127, 2018.

[7] Förster et al. Low-resolution description of the conformational space for intrinsically disordered proteins. *Sci Rep*, 12:19057, 2022.

[8] J. Köfinger, L. S. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer. Efficient Ensemble Refinement by Reweighting. *J Chem Theory Comput*, 15(5):3390–3401, May 2019.

[9] C Lavor, L Liberti, and A Mucherino. The interval Branch-and-Prune algorithm for the discretizable molecular distance geometry problem with inexact distances. *J Glob Optim*, 56:855–871, 2013.

[10] T. E. Malliavin. Tandem domain structure determination based on a systematic enumeration of conformations. *Sci Rep*, 11:16925, 2021.

[11] T. E. Malliavin, A. Mucherino, C. Lavor, and L. Liberti. Systematic Exploration of Protein Conformational Space Using a Distance Geometry Approach. *J Chem Inf Model*, 59(10):4486–4503, 10 2019.

[12] C. J. Oldfield and A. K. Dunker. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*, 83:553–584, 2014.