PhD fellowship subject 2020

Title: **Statistical models and methods for the structural analysis of intrinsically disordered proteins**

Scientific supervisors:
Pierre Neuvial, Institut de Mathématiques de Toulouse (IMT)
Juan Cortés, Laboratoire d'analyse et architecture des systèmes (LAAS-CNRS), Toulouse

Motivation and objectives

This project is motivated by the study of Intrinsically Disordered Proteins (IDPs), a family of proteins that do not fold into a well-defined three-dimensional form. Unlike other proteins, IDPs cannot be represented by a single conformation, and their models must be based on ensembles, usually involving thousands of conformations representing a distribution of states that the protein adopts in solution [1], [2]. In recent years, Juan Cortés (LAAS-CNRS) has collaborated with researchers at the *Centre de Biochimie Structurale* (CBS, Montpellier) on IDP modeling, and they have made original methodological contributions in the framework of the co-supervised interdisciplinary PhD thesis of Alejandro Estaña [3]–[6].
Despite recent progress, IDP modeling remains a largely open research area, offering exciting and challenging problems for the development of new methodological approaches. The goal of this thesis is to develop such novel approaches, based on concepts and recent advances in statistics and artificial intelligence (AI), enabling a more comprehensible connection between amino acid sequences and structural properties of IDPs. A deep understanding of this relationship will enable us to anticipate structural perturbations exerted by sequence modifications and, finally, to design artificial (de-novo) highly flexible proteins with tailored properties, with important applications in health and bio-technologies.

Implementation plan

The thesis project will build on the complementary background in statistics and computer sciences of the two supervisors at IMT and LAAS. In addition, the thesis will involve interactions with biophysicists at *Centre de Biochimie Structurale* (CBS) in Montpelier, experts in IDPs.
The thesis project will involve several tasks, which are strongly interconnected:

- **Construction and statistical analysis of an exhaustive structural database**: This step builds on an existing database of three amino acid fragments (i.e. a tripeptide database) extracted from coil regions in high-resolution experimentally solved protein structures, on which the aforementioned tools are based [3]–[6]. The goal is to construct a more exhaustive and accurate tripeptide database, using a combination of experimental data, extracted from the Protein Data Bank (PDB, https://www.rcsb.org/), and molecular dynamics (MD) simulations. For this, the candidate will have the support of the LAAS and CBS teams. To make the predictive models based on the tripeptide database more accurate and rebuts, a key issue is to properly summarize and represent these conformations as probability distributions (known as Ramachandran distributions) accounting for the local sequence context of each amino acid. Several approaches to this statistical inference task have been proposed, including the use of classical kernel density estimates, or of hierarchical Dirichlet processes [7]. We will apply these classical approaches as well as more recent techniques, based on the Wasserstein distance, to analyze the new database. The strong background in statistics of

Pierre Neuvial will be essential to guide the PhD candidate in this task. Once built and validated, the database and the associated statistical tools (i.e. clustering and testing methods adapted to distributional inputs) will be made publicly available, since they can be of great interest for researchers is structural biology and bioinformatics. The database and the inferred distributions will serve as inputs for subsequent tasks in this thesis.

- **Accurate characterization of conformational ensembles of IDPs**: The distribution of conformations for a given sequence can be modeled as a succession of regions, each of which corresponding to a specific (local) conformation distribution. The core of the thesis project from a statistical perspective is to address two questions outlined below. Both of these questions can be addressed either by aggregating local information into regional information, or performing inference directly at the regional scale.

  Segmentation: The recovery of the location of the breaks between regions from a set of observed conformations, in order to summarize a high-dimensional conformational ensemble into a compact and meaningful description. This question can be cast as a problem of segmentation or multiple change point detection. It raises both statistical challenges (in particular the choice of a number of regions, which is a model selection issue) and computational challenges (the exploration of the space of all possible segmentations) that can be addressed by kernel segmentation approaches [8], [9].

  Tests: The identification of sequence regions that differ between two conformational ensembles (e.g. generated from a reference sequence and a modified sequence). This question can be cast as a problem of test along a sequence. It may be addressed by aggregating local tests of the difference between the distribution of dihedral angles between the two sequences [10], [11], or by constructing measures of (dis)similarity between conformation ensembles that explicitly account for the distributional nature of these ensembles, e.g. Kantorovich-Wasserstein distances for which recent advances in theoretical and computational optimal transport [12], [13] have been made. These metrics will enable a quantitative analysis of the effect of mutations on the structural properties of IDPs, thus helping to understand functional or dysfunctional effects.

- **Experimental evaluation**: For validation of the computational methods developed during the thesis, we will use two disordered proteins, p53 and TIF2, that perform key biological functions and for which our colleagues at the CBS have accumulated a large body of structural and functional information. This validation task will be performed in collaboration with the CBS. Feedback from this evaluation will enable us to improve the computational methods.

Candidate profile

We are looking for a motivated and talented student who should:

- Hold a master's degree in applied mathematics: probability/statistics, machine learning, data science or signal processing.
- Have a strong background in scientific programming, preferably in R and/or Python.
- Have good English skills allowing scientific communication (oral/reading/writing).

Application process

To apply for this position, the candidate is requested to send the following documents in English or French:

- a detailed CV
- a cover letter,
- the marks obtained in Master's degree (of the two years, if available),
- at least one contact reference.

These documents must be emailed to `pierre.neuvial@math.univ-toulouse.fr` and `juan.cortes@laas.fr`.

References

[1] Bernadó, P., L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge, "A structural model for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering", *Proc. Natl. Acad. Sci. USA*, vol. 102, no. 47, pp. 17 002–17 007, 2005.

[2] P. Bernadó and M. Blackledge, "Proteins in dynamic equilibrium", *Nature*, vol. 468, pp. 1046–1048, 2010.

[3] A. Estaña, K. Molloy, M. Vaisset, N. Sibille, T. Simeon, Bernadó, P., and Cortés, J., "Hybrid parallelization of a multi-tree path search algorithm: application to highly-flexible biomolecules", *Parallel Comput.*, vol. 77, pp. 84–100, 2018.

[4] A. Estaña, N. Sibille, E. Delaforge, M. Vaisset, J. Cortés, and P. Bernadó, "Realistic ensemble models of intrinsically disordered proteins using a structure-encoding coil database", *Structure*, vol. 27, no. 2, 381–391.e2, 2019.

[5] A. Estaña, M. Ghallab, Bernadó, P., and Cortés, J., "Investigating the formation of structural elements in proteins using local sequence-dependent information and a heuristic search algorithm", *Molecules*, vol. 24, no. 6, p. 1150, 2019.

[6] A. Estaña, A. Barozet, A. Mouhan, N. Sibille, P. Bernadó, and J. Cortés, "Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments", *Submitted*, 2020.

[7] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. L. Dunbrack Jr, "Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model", *PLOS Comput. Biol.*, vol. 6, no. 4, pp. 1–21, Apr. 2010.

[8] S. Arlot, A. Celisse, and Z. Harchaoui, "A kernel multiple change-point algorithm via model selection", *Journal of Machine Learning Research*, vol. 20, no. 162, pp. 1–56, 2019.

[9] N. Randriamihamison, N. Vialaneix, and P. Neuvial, "Applicability and interpretability of hierarchical agglomerative clustering with or without contiguity constraints", *arXiv preprint arXiv:1909.10923*, 2019.

[10] G. Grazioli, R. W. Martin, and C. T. Butts, "Comparative exploratory analysis of intrinsically disordered protein dynamics using machine learning and network analytic methods", *Frontiers in Molecular Biosciences*, vol. 6, p. 42, 2019.

[11] A. T. Lun and G. K. Smyth, "De novo detection of differentially bound regions for chip-seq data using peaks and windows: controlling error rates correctly", *Nucleic Acids Research*, vol. 42, no. 11, e95–e95, 2014.

[12] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[13] G. Peyré, M. Cuturi, *et al.*, "Computational optimal transport", *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.