2018-00642 - Manifold Learning for Structural Bioinformatics [PhD Campaign]

Level of qualifications required : Graduate degree or equivalent
Fonction : PhD Position

About the research centre or Inria department

Grenoble Rhône-Alpes Research Center groups together a few less than 800 people in 35 research teams and 9 research support departments.

Staff is localized on 5 campuses in Grenoble and Lyon, in close collaboration with labs, research and higher education institutions in Grenoble and Lyon, but also with the economic players in these areas.

Present in the fields of software, high-performance computing, Internet of things, image and data, but also simulation in oceanography and biology, it participates at the best level of international scientific achievements and collaborations in both Europe and the rest of the world.

Context

The main challenge in the field of computational structural biology is to predict and explain molecular flexibility and corresponding conformational changes upon molecular interactions with other molecules, upon evolution or between their functional states.

For example, currently there are no methods that can reliably model structural changes in proteins upon their binding. However, these are crucial to predict the structure of protein complexes with large conformational changes upon association. Such protein complexes are considered as “difficult” in the protein docking community and currently their structure prediction is generally intractable (Vreven et al. 2015). To give another example, flexibility of the protein binding pocket is the major hurdle in reliable prediction of protein-ligand interactions, as it has been demonstrated by a number of blind protein-ligand docking exercises (D3R, GPCR dock, etc.). Finally, intrinsic flexibility of macromolecules is nowadays the limiting factor for high-resolution experimental structure determination (DePristo et al. 2004; Kuzmanic et al. 2014). This is particularly critical for current cryo-EM structure studies (Liao et al. 2015; Jonic 2016), where hardware instrumentation has just evolved to single angstrom resolution (this experimental technique was chosen as Method of the Year 2015 by the journal Nature) (Eisenstein 2016).

Assignment

We have recently developed a novel non-linear dimensionality reduction technique to extract principal non-linear fluctuations of biological molecules (Hoffmann & Grudinin 2017). This will be the starting point of the PhD project. The principal goal of the project will be learning molecular motions from their static structures and applications of this new information.

Main activities

In order to understand how much information on structural dynamics is hidden in static 3D structures, we will start the PhD project with motion extraction from atomic structures of biological molecules. This includes pruning the databases and collecting the training sets, applying linear and non-linear manifold learning techniques to the datasets with many conformations of the same protein, extracting the essential eigenspaces from positional covariance matrices of these structures, and combining first principles harmonic motions with approximate joint diagonalization of the Hessian matrices for the datasets with just a few conformations. More precisely, we will study the quality and applicability of molecular motions computed with joint diagonalization methods and compare them with those directly extracted from positional covariance matrices.

More formally, motion extraction from a set of static molecular structures can be seen as manifold learning (Donoho 2000), a very common technique in machine learning and data analysis. Indeed, even for large macromolecules, such as proteins, the number of their different functional states is very limited (Wei et al. 2016; Keedy et al. 2015; Campbell et al. 2016), and thus we can assume that the motional variability of the structures lie on an embedded manifold within the higher-dimensional space. We can roughly see a manifold as a locally Euclidean space. In computer graphics and vision, for example, two-dimensional manifolds embedded in R3 describe boundary surfaces of 3D objects (Biasotti et al. 2015). More generally, modern methods in data science are founded on the observation that a dataset is a subset of some manifold embedded in a higher-dimensional Euclidean space.
that high dimensional data tend to lie in the vicinity of a low dimensional manifold, thus providing the basis of manifold learning methods (Donoho 2000; Carlsson 2009; Singer & Wu 2011), which generally include two classes, learning linear subspaces, and learning non-linear subspaces.

Finally, we will apply the obtained information for very challenging problems of flexible protein-protein docking and flexible protein-ligand docking. Protein docking is the task of calculating the 3D structure of a protein complex starting from the individual structures of the constituent components. There were many developments in the past on this topic including contributions from myself (Popov & Grudinin 2015; Popov et al. 2014; Neveu et al. 2016). Because proteins have dynamic structures, which may change conformation on binding, predicting how proteins bind is a very challenging task for highly flexible structures and currently no methods can predict such flexible assemblies successfully (Lensink & Wodak 2013; Lensink et al. 2016) despite numerous developed flexible docking methods (Bonvin 2006; Zacharias 2010; Moal & Bates 2010). Here we propose to tackle the problem from the unusual side—first to learn the protein flexibility from the structural knowledge base and then to apply to the flexible docking problem. This will be done using other methods developed in the team (FFT-accelerated sampling, motion-planning techniques, etc).

All the developed algorithms will be rigorously validated on a number of common benchmarks (CASF, D3R, CASP, CAPRI, etc) and distributed to the community. The final methods will be integrated into the SAMSON software platform developed in the team, http://samson-connect.net/. The PhD proposal requires a significant development of source code.

References:


Skills
Excellent oral, written and interpersonal communication skills are essential (working language will be English – knowledge of French is a plus).

Good knowledge of C++ / signal processing / machine learning / orthogonal polynomials / structural chemistry will be an asset.

Benefits package

• Subsidised catering service
• Partially-reimbursed public transport
• Social security
• Paid leave
• Flexible working hours
• Sports facilities

Remuneration

Monthly salary after taxes: around 1596,05€ for 1st and 2nd year. 1678,99€ for 3rd year. (medical insurance included).