INSTITUT DE FRANCE
Académie des sciences

Inria

UNIVERSITÉ CÔTE D'AZUR

VILLE DE NICE

Académie en région
à Nice et Sophia Antipolis

Bio-molecules: on the role of geometry in the triptych structure - dynamics – function

Frédéric Cazals
Université Côte d'Azur, Inria,

Computational Structural Biology (CSB): three main challenges. Computational structural biology ambitions to unveil the relationship between the structure and dynamics of biomolecules (proteins, nucleic acids), and their function. This endeavor is inherently multi-disciplinary: biology and medicine raise questions at the atomic/molecular level, (chemical) physics provides thermodynamic and kinetic models, physics and technology supply instruments giving access to structural and dynamic data, applied mathematics and computer science yield efficient processing methods for these complex and often massive data. This diversity is illustrated by the numerous Nobel prizes awarded over the years, see https://pdb101.rcsb.org/learn/ flyers-posters-and-other-resources/other-resource/structural-biology-and-nobel-prizes, either in Chemistry or Physiology-medicine for Structures and mechanisms (64 up to 2018), or Chemistry or Physics for Methods (11 up to 2018). Particularly relevant is the 2013 Nobel Prize in Chemistry, which was awarded jointly to Martin Karplus, Michael Levitt and Arieh Warshel "*for the development of multiscale models for complex chemical systems.*"

Three main challenges may be identified in CSB:

· **Sequence-structure.** The inference of structures from sequences (amino-acids or nucleotides) is especially important since a mere ~ 105 structures are found in the Protein Data Bank http://www.rcsb.org/), while UniProtKB/TrEMBL (https://www.uniprot.org/) contains of the order of 108 sequences.

· **Structure-dynamics-function.** Biological functions are often coupled to dynamical processes. Unfortunately, massive calculations are required to milestone them by bridging the gap between physically relevant time scales (femtoseconds) and biologically relevant time scales (beyond milliseconds).

· **Large assemblies.** A vast array of biological functions is accomplished by assemblies involving from tens to hundreds of subunits. (The largest assembly of the eukaryotic cell, the nuclear pore complex, comprises circa 500 polypeptide chains.) The inference of the structure and dynamics of such machines requires combining complementary experiments, and raises difficult modeling questions.

Main difficulties. In theory, the aforementioned three challenges can be tackled from first principles i.e. using physical laws. To understand the intrinsic difficulties of such a strategy, it is instrumental to decouple three aspects: structure, thermodynamics, and dynamics. The structure of a macromolecular system requires characterizing active conformations and important intermediates in functional pathways. In assigning occupation probabilities to these conformations, one treats thermodynamics, while transitions between the states correspond to dynamics. These three aspects can be studied using the formalism of potential energy landscape (PEL) [Wal03]. For example, for systems at thermodynamic equilibrium, observables are inherently coupled to density of states (DoS) which intuitively *count* the number of conformations with a given
potential energy. However, because a system with n atoms enjoys $3n$ cartesian coordinates and $d = 3n - 6$ degrees of freedom (upon removing rigid motions), computing DoS requires computing integrals in these high dimensional spaces. Ideally, one would like to undertake such calculations with efficient i.e. polynomial time algorithms, delivering controlled results. However, the development of such algorithms stumbles on intrinsic difficulties of high dimensional spaces i.e. the curse of dimensionality and concentration phenomena. As of now, probabilistic algorithms of the multi-phase Monte Carlo type have only been developed for simpler problems, namely polytope volume calculations [CV16].
Practically, except for highly specialized cases where massive calculations have been used [ea10], neither molecular dynamics (MD) nor Monte Carlo (MC) sampling methods have been able to access the relevant time and length scales for systems of biological interest.
This state of affairs calls for the development of methods exploiting both specific features of the systems scrutinized, and properties of the models used. We now sketch two such methods, which we designed recently based on geometric approaches.

On the importance of geometry: two recent contributions. To focus simulations on regions undergoing large amplitude conformational changes, the first method identifies structurally conserved domains of arbitrary size and possibly non contiguous along the sequence, shared by two conformations of a given molecule (Fig. 1). The method is reminiscent from the bootstrap, as it (i) computes a coarse global structural alignment, (ii) identifies rigid domains from this alignment, using a topological persistence based analysis on a one-parameter family of simplicial complexes, and (iii) bootstraps the alignment employing stable connected components of the aforementioned filtration. The method was success- fully tested on a panel challenging flexible proteins, including in particular fusion proteins. See https://sbl.inria.fr/doc/Structural_motifs-user-manual.html.
The second method addresses the calculation of DoS for peptides and small molecules, using the recently developed stochastic method known as the Wang-Landau (WL) algorithm. The performances of WL rely on the mixing time of the random walk used internally. We designed a

novel random walk exploiting geometric properties, and showed it efficacy to obtain DoS which could not be computed previously.

Software. Software development in CSB is especially challenging due to the interactions between complex biophysical models (coding the physical and chemical properties) and elaborate algorithms (numerical, geometric, topological, combinatorial, statistical). A number of advanced software environments have been developed over the years, for all kinds of data processing and modeling problems. However, these environments do not in general disentangle the end-user applications solving specific biophysical problems, and the underlying low level algorithmic classes. This jeopardizes re-usability and software component optimization. To change the state of affairs, we undertook the design of the Structural Bioinformatics Library (SBL, http://sbl.inria.fr) [CD17], a generic C++/python cross-platform software library targeting complex problems in structural bioinformatics. Its tenet is a modular design offering a rich and versatile framework both of physical models and low level algorithms, which can be combined to develop novel applications without compromising robustness and performances.

Outlook. We noted above that the scarcity of structural data calls for the development of simulation methods delivering accurate predictions for systems ranging in size from one molecule to hundreds of polypeptide chains. While machine learning based methods recently proved efficient in particular for the protein structure prediction problem (see the α-fold approach by DeepMind, at https://deepmind.com/blog/alphafold/), the ability of such models to deliver subtle thermodynamic / kinetic information yet has to be proven.

The tenet of our work is that a deeper understanding of core mathematical / algorithmic questions is needed to change an Art into a Science, and make a stride towards accurate and efficient in silico methods. The importance of such methods cannot be overstated, as a precise understanding of molecular interactions at the atomic/molecular level would open a new era in systems biology (by allowing a precise qualification of interactions in molecular networks), and medicine (via the identification of novel drug targets and the ability to rescue failing functions).

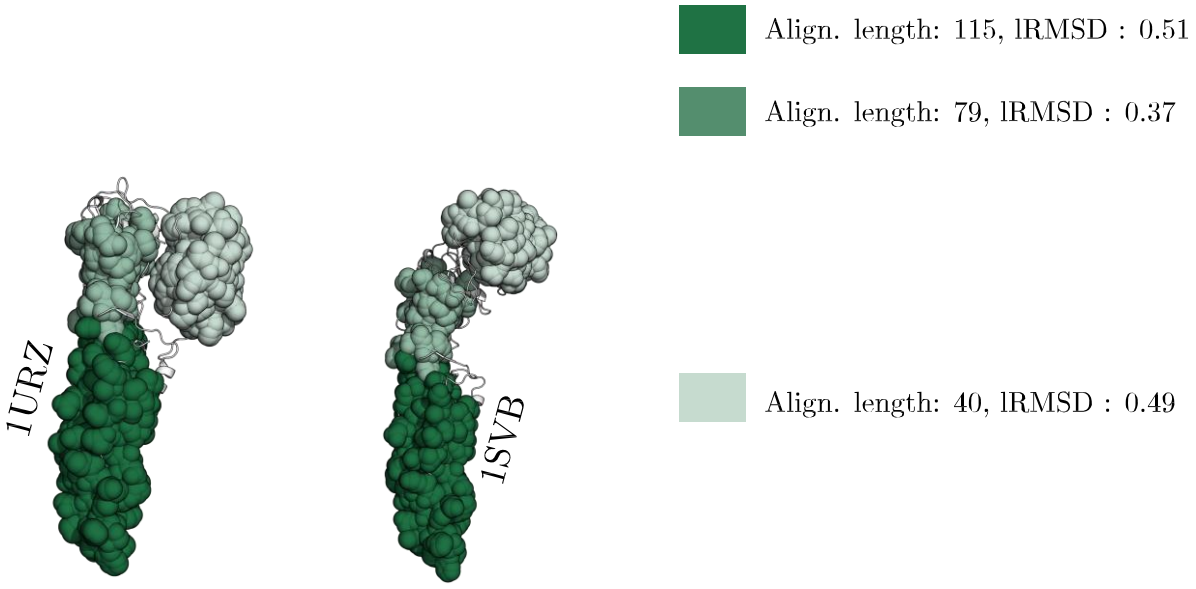

Figure 1: **Three almost rigid domains composing a class II fusion protein undergoing a conformational change between its unbound and bound conformations.** See text for details, as well as https://sbl.inria.fr/doc/Structural_motifs- user-manual.html.

References

[CD17] F. Cazals and T. Dreyfus. The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Bioinformatics*, 7(33):1–8, 2017.

[CV16] B. Cousins and S. Vempala. A practical volume algorithm. *Mathematical Programming Computa- tion*, 8(2):133–160, 2016.

[ea10] D.E. Shaw et al. Atomic-level characterization of the structural dynamics of proteins. *Science*,
330(6002):341–346, 2010.

[Wal03] D. J. Wales. *Energy Landscapes*. Cambridge University Press, 2003.