Post-doctoral position.

Improving inference algorithms for macromolecular structure determination.

Inria Sophia Antipolis - Méditerranée

Supervisors: Frédéric Havet (COATI project-team) and
Dorian Mazauric (ABS project-team).

Abstract.

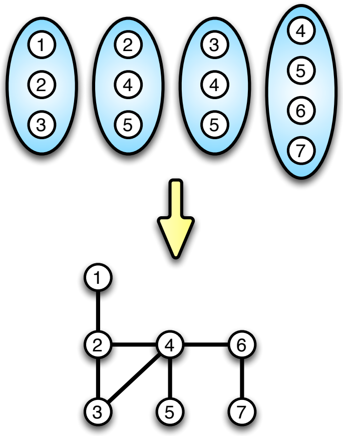
Biological phenomena are based on assemblies of bio-molecules, whose properties depend on structural and dynamic features of their subunits. Experimental studies of such systems face limitations, typically yielding high resolution models of subunits, or low resolution information for the whole assembly. This project focuses on combinatorial algorithms for the connectivity inference problem for mass spectrometry data, so as to discover the interfaces between subunits within an assembly.

Project description

Context. The vast majority of diseases are caused by dysfunction of a protein or set of proteins, and many drugs act on protein active sites to alter their biological function. To address the risks of diseases (prevent, contain, treat, etc), building models of macro-molecular machines is a key endeavor of biophysics because such models offer the possibility to monitor and to fix defaulting systems. Example of such machines are 1) the eukaryotic initiation factors which initiate protein synthesis by the ribosome, 2) the ribosome which performs the synthesis of a polypeptide chain encoded in a messenger RNA derived from a gene, 3) the proteasome which carries out the elimination of damaged or misfolded proteins, etc. These macro-molecular assemblies involve from tens to hundreds of molecules, and range in size from a few tens of Angstroms (the size of one atom) up to 100 nanometers.

But if atomic resolution models of small assemblies are typically obtained with X-ray crystallography and/or nuclear magnetic resonance, large assemblies are not, in general, amenable to such studies. Instead, their reconstruction by data integration requires mixing a panel of complementary experimental data [1]. In particular, information on the hierarchical structure of an assembly, namely its decomposition into sub-complexes (complexes for short in the sequel) which themselves decompose into isolated molecules (proteins or nucleic acids) can be obtained from mass spectrometry. One of the problems, known as connectivity inference problem (CI), aims at finding the most plausible connectivity of the molecules involved.

Model. We assume that the composition, in terms of individual subunits, of selected complexes of the assembly is known [2]. CI problem can be modeled as a combinatorial optimization problem with graphs: a node represents a subunit and there is an edge between two nodes if the two corresponding subunits are in contact in the assembly. *The problem consists in finding a smallest set of edges satisfying some properties on complexes.* For instance, in [3, 4], every selected complex induces a connected graph, the minimal connectivity assumption avoids speculating on the exact (unknown) number of contacts. For example, the top figure represents four complexes and the bottom figure describes an optimal solution composed of seven edges: every set of nodes of the four complexes induces a connected subgraph. The problem has also been studied in different contexts and for different variants (e.g. every complex induces an induced tree/path/star) [5]. The main objectives concern the design of efficient algorithms for general models.



Research program. Our goal is to design efficient algorithms for A) the problems previously described with general properties, B) for the associated problems of new models taking into account multiples copies of the subunits, and C) for the problems related to enumeration and comparison of different solutions.

A) First, we plan to develop algorithms handling combinatorial constraints reflecting biophysical properties: bounded maximum degree (a protein has a limited number of neighbors), (induced) subgraph contained in a complex (when selected contacts are already known or symmetry constraints), constraints on the diameter, etc. For each variant of the problem, the first aim is to determine the complexity of the problem (polynomial-time solvable, NP-hard). In (the very likely) case the problem is NP-hard, the goal is then to develop efficient approximation algorithms or to prove that this problem is hard to approximate (APX-hard). We also plan to develop parameterized algorithms and/or moderately exponential algorithms. The same study for different instance classes is also planned in order to obtain faster and/or more accurate algorithms for specific problems (e.g. by integrating biophysical assumptions).

B) Secondly, we aim at proposing new generalized models taking into consideration multiple copies of the subunits – a situation typically faced for macro-molecular systems with symmetries. The first issue consists in formalizing a suitable new mathematical model because of the limitations of the original one to take into account such specific properties. We plan to develop efficient algorithms (approximation, parameterized, moderately exponential) for these different problems. For instance, we aim at designing reduction rules in order to reduce the size of the instance and develop efficient branch-and-bound algorithms.

C) Thirdly, the number of minimal solutions can be large and some solutions with few more edges may be interesting. Thus, we plan to study enumeration problems in order to find large sets of candidate solutions. We then model the properties of good solutions and develop efficient algorithms to find them. For instance, it is possible to quantify the difference between these candidate solutions and to improve existing works on the computation of so-called consensus solutions [3, 4].

Finally, the methods developed will be tested on classical systems (proteasome, eukaryotic initiation factors, nuclear pore complex), and also on systems currently investigated by collaborators of the ABS project-team (in particular Perdita Barran, Manchester University).

Références

[1] F. Alber, F. Forster, D. Korkin, M. Topf, and A. Sali. Integrating diverse data for structure determination of macromolecular assemblies. *Ann. Rev. Biochem.*, 77:11.1–11.35, 2008.

[2] T. Taverner, H. Hernández, M. Sharon, B.T. Ruotolo, D. Matak-Vinkovic, D. Devos, R.B. Russell, and C.V. Robinson. Subunit architecture of intact protein complexes from mass spectrometry and homology modeling. *Accounts of chemical research*, 41(5):617–627, 2008.

[3] D. Agarwal, J.-C. S. Araujo, C. Caillouet, F. Cazals, D. Coudert, and S. Pérennes. Connectivity inference in mass spectrometry based structure determination. In *ESA 2013*, volume 8125 of *LNCS*, pages 289–300. 2013.

[4] D. Agarwal, C. Caillouet, D. Coudert, and F. Cazals. Unveiling contacts within macro-molecular assemblies by solving minimum weight connectivity inference problems. *Molecular and Cellular Proteomics*, 14:2274–2282, 2015.

[5] I. D. Mantas. The subset interconnection design problem: algorithms and special cases. Master thesis, 2015.

[6] K. Lasker, M. Topf, A. Sali, and H.J. Wolfson. Inferential optimization for simultaneous fitting of multiple components into a cryoem map of their assembly. *Journal of molecular biology*, 388(1):180–194, 2009.

[7] F. Alber, S. Dokudovskaya, L. M. Veenhoff, W. Zhang, J. Kipper, D. Devos, A. Suprapto, O. Karni-Schmidt, R. Williams, B.T. Chait, A. Sali, and M.P. Rout. The Molecular Architecture of the Nuclear Pore Complex. *Nature*, 450(7170):695–701, Nov 2007.

[8] T. Dreyfus, V. Doye, and F. Cazals. Assessing the reconstruction of macro-molecular assemblies with toleranced models. *Proteins: structure, function, and bioinformatics*, 80(9):2125–2136, 2012.

[9] T. Dreyfus, V. Doye, and F. Cazals. Probing a continuum of macro-molecular assembly models with graph templates of sub-complexes. *Proteins: structure, function, and bioinformatics*, 81(11):2034–2044, 2013.