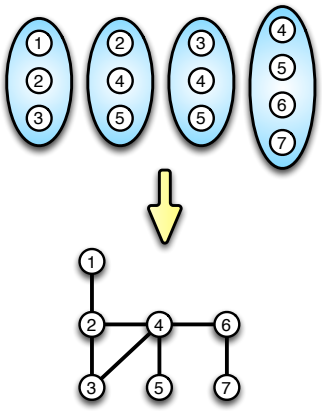
PhD thesis proposal

Laboratory: Inria Sophia Antipolis - Méditerranée

Supervisors: Frederic.Havet@cnrs.fr (I3S et Inria, COATI, `https://team.inria.fr/coati`)
Dorian.Mazauric@inria.fr (Inria, ABS, `http://team.inria.fr/abs`)
In close collaboration with Rémi Watrigant (LIP, ENS Lyon & Université Claude Bernard Lyon 1, `www.ens-lyon.fr/LIP/MC2`).

Graph Algorithms techniques for (low and high) resolution models of large protein assemblies

Graph Problems. Two main problems will be addressed in the thesis. The first one is the *Minimum Connectivity Inference* (MCI) problem. Let H be a hypergraph with $V(H)$ its set of vertices and $E(H)$ its set of hyperedges. The MCI problem consists in finding a smallest set of edges E satisfying the following constraint: the set of nodes of every hyperedge of H must induce a connected subgraph in $G = (V(H), E)$. For example, the top figure represents the four hyperedges of an instance and the bottom figure describes an optimal solution composed of seven edges: every hyperedge induces a connected subgraph [AAC+13, ACCC15]. See the bioinformatics context below.
The second one is the so-called *domino* problem. Let $G = (V, E)$ be a graph and let $C(v)$ be a list of possible configurations for every node $v \in V$. The domino problem consists in computing a configuration for each node maximizing some objective functions. Again, see the bioinformatics context below.



Research programme. The aim of this PhD thesis is to develop algorithms for some generalized versions of both problems handling combinatorial constraints reflecting biophysical properties (bounded maximum degree, constraints on the diameter of the graph sought, ...). 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 problems are 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).

Context. A macromolecular assembly is composed of subunits (e.g. proteins or nucleic acids). We assume that the composition, in terms of individual subunits, of selected complexes of the assembly is known. Indeed, a given assembly can be chemically split into

complexes by manipulating chemical conditions, and the composition of these complexes can then be inferred using native mass spectrometry. A node represents a subunit, while the hyperedges represent the different complexes. The MCI problem consists in inferring the *contact graph* of these subunits, where an edge between two nodes means that the two corresponding subunits are in contact in the assembly. Hence, *the MCI problem consists in finding a smallest set of contacts satisfying the connectivity constraints on complexes.* In a second time, this contact graph will define the input graph for the domino problem that consists in determining the high resolution structure of a given assembly. A configuration of a node is so a conformation of the corresponding protein (that is a position of each of its atoms in $\mathbb{R}^3$) and is obtained by X-ray cristallography. This process involves two mains steps, namely computing conformations of subunits compatible with low resolution data [CDM$^+$15, RDRC16], and assembling these conformations to build the assembly [ACW15]. Both steps can be phrased as optimization and enumeration problems involving graphs. See also [AFK$^+$08, THS$^+$08].

To summarize, the goals of the two graph problems considered in this thesis are the determination of low resolution structure of given assemblies and the reconstruction of atomic resolution models of the large protein assemblies.

Background. Theoretical computer science and/or bioinformatics and/or applied mathematics.

References

[AAC$^+$13] D. Agarwal, J. Araujo, C. Caillouet, F. Cazals, D. Coudert, and S. Pérennes. Connectivity inference in mass spectrometry based structure determination. In H.L. Bodlaender and G.F. Italiano, editors, *European Symposium on Algorithms (Springer LNCS 8125)*, pages 289–300, Sophia Antipolis, France, 2013. Springer.

[ACCC15] 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.

[ACW15] N. Amir, D. Cohen, and H. Wolfson. Dockstar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes. *Bioinformatics*, 31(17):2801–2807, 2015.

[AFK$^+$08] F. Alber, F. Förster, 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.

[CDM$^+$15] F. Cazals, T. Dreyfus, D. Mazauric, A. Roth, and C.H. Robert. Conformational ensembles and sampled energy landscapes: Analysis and comparison. *J. Comp. Chem.*, 36(16):1213–1231, 2015.

[RDRC16] A. Roth, T. Dreyfus, C.H. Robert, and F. Cazals. Hybridizing rapidly growing random trees and basin hopping yields an improved exploration of energy landscapes. *J. Comp. Chem.*, 37(8):739–752, 2016.

[THS$^+$08] 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.