Master internship proposal: Connectivity inference in structural proteomics

Context and multi-disciplinary nature.  Molecular assemblies involving from tens to hundreds of macro-molecules (proteins or nucleic acids) are commonplace, yet, very little is known on their structure. To fill this gap, a number of biophysical experiments (mass spectrometry, tandem affinity purification, etc.) providing information on the composition of sub-systems called oligomers are being developed \cite{SR07}. More precisely, given an assembly, one such experiment gives access to the composition of one sub-system, called oligomer, so that experiments for multiple overlapping oligomers convey information on the overall connectivity. For example, if one oligomer reduces to two proteins, these proteins touch in the assembly. Connectivity inference is the problem concerned with the elucidation of this connectivity. Instances for the connectivity inference problem can be modeled as a graph in which each vertex corresponds to a protein, and an assembly (oligomer or set of proteins) is a connected subset of vertices.

The problem is of multi-disciplinary nature, since developing solutions requires a virtuous circle between bio-physics on the one hand, and algorithmic - optimization - graph theory on the other hand.

Goals.  The connectivity inference problem in the particular case where each protein species is represented by a single copy has recently been studied using tools from graph theory and integer linear programming \cite{AAC13, ACCC}. Roughly, the minimum connectivity inference problem is to find the smallest subset of edges of a complete graph such that each subset of vertices associated to an assembly gets connected. This optimization problem has been proved NP-hard and hard to approximate in general. Nonetheless, when the number of proteins is moderate (less than 20), the problem can be solved efficiently using mixed integer linear programming (MILP) and a greedy approximation algorithm.

These two types of algorithms generate ensembles of (optimal) solutions, from which consensus solutions were defined. Tests conducted on systems with up to 20 proteins show an almost perfect agreement between the predicted contacts and the experimentally determined ones.

The goal of this internship is to go beyond this state of the art, in the following directions:

- Understand the relationship between the structure of the oligomers and the reported solutions.
- Exploit hypothesis on the geometry of the proteins/oligomers, to speed up the optimization process.
- Generalize the modeling in order to handle the cases where multiples copies of a protein are present.

Misc.  Remuneration according to the Inria grid. Duration: 6 months; possibility to follow-up with a PhD thesis.

References.


\cite{ACCC} D. Agarwal, C. Caillouet, F. Cazals, and D. Coudert. Unveiling contacts within macro-molecular assemblies by solving minimum weight connectivity inference problems. submitted.
