Machine learning and guided docking for protein-protein interactions using massive genomic data

Mots clés : Array

Directeur de thèse : Martin Weigt
Co-encadrant(s) :
Unité de recherche : Laboratoire de Biologie Computationnelle et Quantitative
École doctorale : École Doctorale Informatique, Télécommunications, Electronique de Paris
Domaine scientifique principal : Sciences et technologies de l'information et de la communication

Résumé du projet de recherche (Langue 1)

Proteins are the major workhorses of the cell. However, few proteins exert their function in isolation. Rather, most proteins take part in concerted physical interactions with other proteins, forming complex networks of protein-protein interactions (PPI). Unveiling the PPI organization at different biological scales is one of the most formidable tasks in biology today.

The experimental characterization of PPI is far from being satisfying. Computational approaches emerge as an attractive alternative. Our project aims at bringing together two complementary approaches, developed by the scientific community over the last years: (i) Molecular docking starts from individual protein structures (possibly modeled themselves), and tries to computationally assemble their complexes using detailed microscopic models. Docking frequently predicts an enormous multitude of alternative structure models; the identification of the correct one based only upon the knowledge of the monomeric proteins remains a hard and error-prone task. Our collaborator Raphaël Guérois (CEA Saclay) is one of the leading experts of these techniques. (ii) Alternatively, coevolutionary modeling builds upon increasingly abundant genomic databases, which provide large samples functionally en evolutionary related proteins. In this context, the proponent Martin Weigt has developed the Direct Coupling Analysis (DCA) to predict inter-protein residue-residue contacts from statistical modeling of evolutionary related sequences. While DCA has shown high utility to guide protein docking in a number of example cases, the broad applicability is currently hindered by two factors – (i) coevolutionary modeling requires large multiple-sequence alignments (MSA of ~1000 sequences, frequently not yet available for interacting proteins), and (ii) the unsupervised nature of DCA limits the accurate association of coevolutionary signals with residue-residue contacts.

In the related but simpler case of tertiary protein-structure prediction, recent breakthrough has been achieved using known protein structures to develop supervised machine-learning techniques for the coevolutionary contact prediction. RaptorX, the currently most accurate predictor (winner of the last edition of the CASP [Critical Assessment of protein Structure Prediction, http://predictioncenter.org/] competition), uses deep convolutional networks combining input from coevolutionary analysis with sequence profile features, predicted secondary structure and solvent accessibility, to obtain highly accurate predictions even in medium-size MSA (~100 sequences), where unsupervised DCA-type techniques fail.

The project aims at making similar progress in the harder and more important case of protein-complex assembly, using a combination of data-driven approaches from machine learning and statistical inference, and computational modeling of biomolecular complexes via docking.

Specific aims of the project are: (1) to build a large gold-standard database of interacting proteins for training and test purposes; (2) to critically benchmark coevolutionary modeling and its capacity to predict inter-protein contacts; (3) to develop highly accurate supervised approaches for contact prediction using deep learning; (4) to use predicted contacts to guide the computational de novo assembly of protein complexes. These tasks are part of the long-term effort of the team to develop data-driven computational approaches to connect all scales of PPI – from entire genomes down to individual residues – towards an evolutionary informed structural systems biology.

Informations complémentaires (Langue 1)
The project relies on collaborations in France (Raphael Guerois, CEA) and abroad (Alexander Schug, Karlsruhe, Germany; Hendrik Szurmant, Los Angeles, USA; Andrea Pagnani, Turin, Italy). The candidate will fully benefit from this collaboration, e.g. via research visits in the French and European partner institutes.

Key references:

