Proposition de recherche doctorale

From residue co-evolution to protein structure prediction

Mots clés :
- Directeur de thèse : Martin Weigt
- Co-encadrant(s) :
- Unité de recherche : Laboratoire de Biologie Computationnelle et Quantitative
- Ecole doctorale : Ecole Doctorale Informatique, Télécommunications, Électronique de Paris
- Domaine scientifique principal: Divers

Résumé du projet de recherche (Langue 1)

The grand challenge of biology in the 21st century, to become a quantitative science like physics and chemistry, can be solved only by integrating recent progress in computer science (high-performance computing as well as novel methods in statistical inference and model learning) and experimental biology (high-throughput sequencing). Within this project, our vision is to develop a powerful algorithmic framework whereby the mining of vast amounts of raw data will lead to the understanding of complex biological processes. More specifically, we will exploit the sequence variability of related proteins across thousands of sequenced genomes, to detect evolutionary constraints, and to exploit them for the prediction of protein structures (contact map and 3D fold prediction). Protein-structure prediction is recognized as one of the most important problems in bioinformatics, medicine and biotechnology. In fact, in the course of evolution, protein structure is remarkably conserved, whereas amino-acid sequences vary strongly between homologous, i.e. evolutionarily related proteins (so-called protein families). This structural conservation constrains sequence variability, forcing residues to co-evolve: residues being close in the protein structure (but possibly distant along the sequence) will typically evolve in a correlated way, cf. Valencia et al. 2013 for a review. In our team, we have recently proposed an innovative statistical inference method, called Direct-Coupling Analysis (DCA), which turned out to reach a substantial breakthrough in detecting residue-residue contacts from sequence information alone (Weigt et al. 2009, Morcos et al. 2011, Ekeberg et al. 2013). This inference approach is based on the statistical modeling of protein sequences by Markov Random Fields (MRF). This problem, being a priori infeasible for protein data due to its exponential time complexity, has been approached with methods inspired my statistical physics (mean-field approximation) and machine learning (pseudo-likelihood maximization).

Résumé du projet de recherche (Langue 2)

The accuracy in contact prediction remains limited by the fact that inference by DCA is completely blind with respect to any form of structure underlying the contact prediction problem. This structure can roughly be viewed as the set of bio-chemical constraints representing our general knowledge about proteins (the same "rules" that draw the thin line between all the possible sparse matrices and the subset of sparse matrices representing contact maps with real biological meaning) and thus can be abstracted by reconducting the contact prediction problem to some other well-known structured problems in machine learning, such as image and object recognition. This PhD projects aims at a substantial increase in accuracy of contact map prediction by the integration of such information into DCA. To this aim, we will undertake three major steps: Phase 1: Large-scale analysis of the co-evolutionary signal in large protein families The most important protein-family database, named Pfam, currently lists almost 15,000 protein families, out of which more than 3,000 contain at least 1,000 sequences (providing enough statistics for robustly learning the parameters of the MRFs) and provide known 3D protein structures (thus allowing for evaluating the predictions). We will apply DCA to these families, and analyze the detected signal in terms of a number of protein sequence annotations (amino-acid properties, secondary structure etc.). This will allow us to understand which kind of contacts show detectable co-evolutionary information. Phase 2: Integration of sequence annotations into inference by DCA We aim at using the findings of Phase 1 to improve the contact-map prediction of DCA substantially beyond its current level. To this aim, we want to integrate supplementary information into DCA. We will follow two main strategies: (i) posterior filtering of DCA scores with sequence annotations, to achieve a classification of DCA predictions into potential true-positive and false-positive predictions; (ii) usage of sequence information in form of prior knowledge in the MRF parameter learning done by DCA. Both require the application of advanced learning techniques. Phase 3: Exploitation of improved DCA for protein contact prediction In between the above-mentioned 15,000 protein families in Pfam, there are slightly more than 1,000 which contain enough sequence information for statistical analysis, but for which no 3D protein structure is known. We will use the algorithmic developments achieved in Phase 2 to obtain contact map predictions for these families. As a last and final step, we will use techniques like the one proposed in (Marks et al. 2012, Sulkowska et al. 2012) to actually predict the 3D structures for these 1,000 families.

Informations complémentaires (Langue 1)

The candidate will benefit from tight international collaboration of the hosting team with the groups of A. Schug (Karlsruhe Institute of Technology, Germany), H. Szurmant (Scripps Research California, USA), A. Pagnani (Politecnico Turin, Italy), J. Sulkowska (Warsaw University, Poland) and national collaborations with R. Monasson and S. Cocco (ENS Paris) and O. Tenaillon (INSERM).