Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Baldo Oliva
Email	baldo.oliva@upf.edu
Institution	UPF
Website	sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

Protein Folding ab initio, based on contact maps and supersecondary structures using distance-restraints derived from multiple sequence alignments.

Keywords:

ab initio fold prediction; protein design; threading

Summary:

Background We work on the modeling of macro-complex regulatory structures formed by transcription factors (TF), co-factors, and DNA. One of the main contingencies to structurally model the macro-molecular complex formed by TFs and co-factors either in the enhancer or promoter binding sites is the completion of the structures. Modelling the structure of the DNA-binding domain of a TF bound to DNA can be achieved by homology modelling. However, the domains of TF that bind co-factors or other elements of the macro-molecular complex often involve unstructured or non-structurally solved regions. To solve this contingency we propose two approaches: 1) we will use docking of binary interactions if the structure of the unbound proteins is known; and 2) if the structure of one or both partners of a binary interaction we will apply threading and ab initio approaches to generate a structural model. Both approaches can be addressed by standard methods(1,2), but here we propose to update two innovations produced in our group during the previous grant. First, we developed iFrag to predict binding interfaces of proteins using only the sequences of the partners(3); second, we developed RADI (https://www.biorxiv.org/content/10.1101/406603v1), a program to predict residue-residue interactions based on amino-acid covariations in the sequence using Direct Coupling Analysis. We plan to combine RADI and iFrag to improve the prediction of binding and apply the approach on directed docking, modelling the structure of binary interactions when the structures of both partners are known or modelled. Furthermore, we will use RADI, combined with local super-secondary structures, named sMotifs, as classified in ArchDB(4), to model ab initio the structure of proteins that cannot be modelled by homology modelling. Approach We have already developed a pilot version, using restraints obtained with RADI, the prediction of secondary structure, and short fragment sMotifs templates from ArchDB to apply MODELLER(5) and construct an ab initio model structure of a query sequence. However, we need to update this version to be applicable on a largest extend of proteins. First, we need to update the database of sMotifs. Second, many proteins can share similar local conformations but still they may not be classified. Therefore, we will extend the number of short-fragment templates by searching with psi-blast and selecting all short fragments with known structure that can potentially be used as templates. One of the problems already detected in the pilot is that more than one template can be assigned to the same region of the query sequence. We selected the longest fragment in the original version of the pilot, but the most appropriate approach is to model all combinations of fragments to produce several models after the application of restraints. Then, because neither the restraints of RADI are exact, nor the short-fragment templates correct, we will have several but different solutions of the conformation, so we will cluster the solutions, score and rank the conformers by statistical potentials using SPserver (6) and select the common restraints and common short fragment templates. The scores of SPserver can be improved by training a combination of energies that benefits from a large knowledge of structures and the use of several potentials. Also, we need a machine-learning method to optimize the selection of structures. A potential approach is to split the alignment in sections of 100 column-positions by sliding windows and construct the final model from the overlaps. The structure of the model will be iteratively repeated and improved until all convenient restraints and templates are congruent with the model.

References:

1. Garcia-Garcia, J., Bonet, J., Guney, E., Fornes, O., Planas, J. and Oliva, B. (2012) Networks of ProteinProtein Interactions: From Uncertainty to Molecular Details.

Mol Inform, 31, 342-362. 2. Mirela-Bota, P., Aguirre-Plans, J., Meseguer, A., Galletti, C., Segura, J., Planas-Iglesias, J., Garcia-Garcia, J., Guney, E., Oliva, B. and Fernandez-Fuentes, N. (2020) Galaxy InteractoMIX: An Integrated Computational Platform for the Study of Protein-Protein Interaction Data. J Mol Biol. 3. Garcia-Garcia, J., Valls-Comamala, V., Guney, E., Andreu, D., Munoz, F.J., Fernandez-Fuentes, N. and Oliva, B. (2017) iFrag: A Protein-Protein Interface Prediction Server Based on Sequence Fragments. J Mol Biol, 429, 382-389. 4. Bonet, J., Planas-Iglesias, J., Garcia-Garcia, J., Marin-Lopez, M.A., Fernandez-Fuentes, N. and Oliva, B. (2014) ArchDB 2014: structural classification of loops in proteins. Nucleic Acids Res, 42, D315-319. 5. Webb, B. and Sali, A. (2016) Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics, 54, 5 6 1-5 6 37. 6. Aguirre-Plans, J., Meseguer, A., Molina-Fernández, R., Marin-Lopez, M.A., Jumde, G., Casanova, K., Bonet, J., Fornes, O., Fernandez-Fuentes, N. and Oliva, B. (2020) SPServer: Split-Statistical Potentials for the analysis of protein structures and protein-protein interactions. BMC Bioinformatics.

Expected skills::

Python programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed