upf. 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	http://sbi.upf.edu
Group	SBI

Project

Structural bioinformatics

Project Title:

Proposal to score protein structures using distance-restraints derived from multiple sequence alignments and statistic potentials.

Keywords:

Protein structure quality

Summary:

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 have used 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. 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 constructed an ab initio model structure of a query sequence. However, we need to update this version to select the correct folds, because neither the restraints of RADI are exact, nor the short-fragment templates correct, and as a consequence, we obtain several but different solutions of the conformation. Therefore, we must cluster the solutions, score and rank the conformers and select the best conditions of the models. We have developed a program of statistical potentials, the SPserver (6) to select the correct folds. 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. The program can also benefit from the results of RADI to improve the learning. This machine-learning method will be useful to optimize the selection of structures. Proposed approach. 1) Define a method to compare as residue-profile the structures of model decoys with the experimental structure. This can be obtained by the superposition of the complete structure, using the RMSD fluctuation of CA atoms, the local matrix of distances around residues, or the local GDT score. Local distance matrices have the benefit to compare straightforward with potentials under a distance cut-off threshold. The information will be transformed into a score with values between 0 and 1, being 1 the best match (for example, the ratio of identic contacts within a radius that are the same between the decoy and the experimental structure). 2) Construct several ML methods of python “sklearn” library (SVM, NN, LogisticProgression, RF, etc.) using the statistic potential profiles per residue as inputs, the matrix of interacting potentials with the residues within a 3D-ball and the RADI direct and mutual information. The output will be a normalized score between 0 and 1 of the local quality. Additional approaches, using PyTorch, Keras and Tensorflow will also be considered. 3) The final score of the model will be obtained as the average of the total of scores along the sequence.

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