Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Group

Project

Structural bioinformatics

Project Title:

Unraveling signaling bias at G protein-coupled receptors (GPCRs)

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G protein-coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. GPCRs can initiate signalling through binding with several intracellular partners, which in turn elicit distinct signalling cascades. It is established that in physiological conditions, a receptor binds to each of the partners in equal proportion. Interestingly some drugs act by altering the GPCR structure so that it preferentially binds to one specific partner - a phenomenon known as signaling bias. Such ligands, named biased ligands, offer great promise, as they enable to modify pathways associated with symptoms while not modifying other pathways - which could cause side effects. However, the molecular requirements for a molecule to act as a biased agonist within a receptor are still poorly understood. Molecular dynamics (MD) is a novel and sophisticated technique that enables to simulate protein behaviour in a physiological environment. The Master student will apply this approach to study time-resolved molecular mechanisms underlying signaling bias induced by small drug-like molecules. For this, the student will be trained on setting up simulated systems, running production simulations as well as the application of a wide range of analysis tools that allow capturing subtle structural events related to signaling bias. Structural insights will be exploited for the design of a novel class of GPCR modulators with a tailored signaling profile. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

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

upf. Universitat Pompeu Fabra *Barcelona*

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu

Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Lab

Project

Structural bioinformatics

Project Title:

Why is water essential for GPCR functionality?

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G protein-coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. As for the whole life, water is critical for GPCR functionality at a nanoscopic scale. It forms important intermolecular networks that mediate the signaling response of the receptor. Our group owns an extraordinary dataset with unprecedented information about the implication of water molecules in GPCR structural dynamics and the binding of small drug-like molecules. We believe that stabilized or disrupted intermolecular water signatures drive the functional consequences of a drug-like molecule. In this project, the Master student will seek for the underlying molecular mechanism of the water-mediated functional responses. For this, she/he will (i) develop an analysis pipeline to extract relevant information from our unique dataset and (ii) setup control simulations to validate obtained conclusions. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Lab

Project

Structural bioinformatics

Project Title:

Simulation meets experiment: refinement of cryo-EM GPCR structures using molecular dynamics

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G-protein coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. Recent advances in cryo-electron microscopy (cryo-EM) and image classification provide insights into ensembles of low- to high-resolution that describe differently populated conformational states of proteins. However, methods for deriving accurate atomistic models from cryo-EM density maps lag behind this resolution revolution. The increasing amount of molecular detail requires the development of new methodologies and software to accurately and timely interpret experimental densities. Molecular dynamics (MD)-based refinement methods have grown into a valuable approach to tackle this challenge. In this project, the Master student will develop an MD-based pipeline that can be applied to GPCRs. This represents an important milestone for the scientific community as it can provide novel structural insights into this important drug targeting class. For this, the student will learn how to setup simulations and how to use correlation-driven MD for the refinement of atomistic models into cryo-electron microscopy maps. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y Igaev et al. Automated cryo-EM structure refinement using correlation-driven molecular dynamics, Elife 2019, DOI: 10.7554/eLife.43542

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes

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

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:

TF-DNA Binding strength affected by methylations

Keywords:

enhancer methylation

Summary:

Changes in DNA methylation are involved in development, disease, and the response to environmental conditions. Methylation of DNA is thought to regulate transcription both directly and indirectly. CpG methylation can directly repress transcription by preventing binding of some transcription factors (TFs) to their recognition motifs(17). For further insights, Lea et al. developed mSTARR-seq(18), a method that assesses the causal effects of DNA methylation on regulatory activity at genomic high-throughput level. Our objective is to predict the changes of TF binding caused by methylation. First we will build a database of methylated DNA binding with known TF binding. In a first approach the database will be extracted from experimental data of Yin et al. (17) and Lea et al. (18), indicating the loss or gain of TF binding. In a second approach, we will infer the effect from the comparison of bound TF binding sites with and without methylations. We will use the dataset of UniBind(13) to select the binding sites confirmed bound by TFs or the predictions of Viestra et al. (15). Then, we will select the tracks from UCSC Genome Browser with assays of DNA methylation (i.e. Methyl-RBBS) specific for tissue. We will compare the percentage that cytosines are methylated in the binding site with respect to any other location in the genome (this can be further refined by comparing with cytosines in the same TAD region). We will use the hypergeometric distribution to compare the ratio of methylation versus the expected ratio according to the length of binding recognition (as derived by the ChIP-Seq experiment). We will split the results in three categories: 1) If the ratio of methylation is lower than expected, then the methylation of cytosines reduces the TF binding. 2) On the contrary, if the methylation in the binding regions is higher than expected, the methylation is required for TF binding. 3) Otherwise, the methylation has no effect on TF binding. With the new bindings (case 2) we will generate statistical potentials specific of methylated cytosines by modelling the structure of TF-DNA binding, introducing a new symbol for methylated cytosines and including them in the general statistical potentials. We will calculate statistical potentials specific of the family of each TF including the new symbol for methyl-cytosine as in Meseguer et al. (19). The effect of disruption (case 1) will be used to generate statistical potentials specific of disruption. As before, the structure of TF-DNA binding will be modelled and the frequencies of the interactions between amino-acids and nucleotides will be obtained from the models. However, these potentials will be used to determine the potential of disruption, as these are the models of interactions lost after methylation. As before, a general potential will be derived with all TFs and their disrupted DNA binding sites and another set of potentials, specific for each TF family will be constructed. Finally, we will test the capacity of predicting TF disruptions after cytosine methylation or TF-DNA new bindings and specific PWMs for methyl cytosines. Two tests will be used for validation. First, using a 5-fold protocol with partially hidden data; and second, by training the method with one set of methylation (i.e. using the experiments of Yin et al. (17) and Lea et al. (18)) and testing the potentials in a different set (i.e. using data of ENCODE and removing redundancies with the training).

References:

1. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. Nucleic Acids Res, 43, W39-49. 2. Fuxman Bass, J.I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M. and Walhout, A.J. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. Cell, 161, 661-673. 3. Fuxman Bass, J.I., Pons, C., Kozlowski, L., Reece-Hoyes, J.S., Shrestha, S., Holdorf, A.D., Mori, A., Myers, C.L. and Walhout, A.J. (2016) A gene-centered C. elegans protein-DNA interaction network provides a framework for functional predictions. Mol Syst Biol, 12, 884. 4. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327-339. 5. Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr. and Reddy, T.E. (2015) Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res, 25, 1206-1214. 6. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S. and Kellis, M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol, 34, 1180-1190. 7. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D. and Shendure, J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol, 27, 1173-1175. 8. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B. et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol, 30, 271-277. 9. Fornes, O., Gheorghe, M., Richmond, P.A., Arenillas, D.J., Wasserman, W.W. and Mathelier, A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. Sci Data, 5, 180141. 10. Kumar, S., Ambrosini, G. and Bucher, P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res, 45, D139-D144. 11. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A. et al. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583, 699-710. 12. Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D. et al. (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res, 41, D171-176. 13. Gheorghe, M., Sandve, G.K., Khan, A., Cheneby, J., Ballester, B. and Mathelier, A. (2019) A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res, 47, e21. 14. Mathelier, A., Shi, W. and Wasserman, W.W. (2015) Identification of altered cis-regulatory elements in human disease. Trends Genet, 31, 67-76. 15. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E. et al. (2020) Global reference mapping of human transcription factor footprints. Nature, 583, 729-736. 16. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A. et al. (2020) Index and biological spectrum of human DNase I hypersensitive sites. Nature, 584, 244-251. 17. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F. et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science, 356. 18. Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E. and Tung, J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. Elife, 7. 19. Meseguer, A., Årman, F., Fornes, O., Molina-Fernández,

R., Bonet, J., Fernandez-Fuentes, N. and Oliva, B. (2020) On the prediction of DNA-binding preferences of C2H2-ZF domains using structural models: application on human CTCF. NAR Genomics and Bioinformatics, 2.

Expected skills::

Python programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

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:

TF-DNA Binding strength affected by mutations on DNA binding site

Keywords:

Single nucleotide variants; Cis-regulation; Transcription factors;

Summary:

The principal reason to understand changes on the binding strength of a TF for a specific DNA binding sequence is to study cis-regulation and mutations that will affect it. Therefore, we focus on the prediction of the effect of one or more single nucleotide substitutions disrupting the recognition of the specific TF. The approach can be used to predict causal regulatory haplotypes that likely contribute to human phenotypes and to functionally fine map causal regulatory variants in regions of high linkage disequilibrium identified by expression quantitative trait loci (eQTL) analyses. We will predict binding strength to classify strong and weak changes and deciding if they imply the loss of the interaction. We will apply structural modelling with several templates, testing all potential PWMs, and the analysis of various statistical potentials on the TF-DNA interaction, comparing the native binding site with all other DNA variants. In order to train an AI/ML model, we will use the information from experimental Protein Binding Micro-arrays (PBM) of each TF. We will use as inputs the profiles of statisticalenergies and also the profiles of the PWM search along the DNA sequence, using the scores of FIMO (1) and the enrichment achieved with different structural models of the same TF. In order to learn and test with these scores, using different TFs and DNA binding lengths, the scores of the profiles will be normalized. The normalization will help us to better characterize the magnitude of the change produced by single nucleotide substitutions of the DNA binding sequence. The use of PBM will help us to handle an overwhelming amount of information, as all combinations of DNA sequences (formed by 8 or 12-mer nucleotides) are experimentally tested. We will train on PBM and the test will be performed on PBM (using a 10-fold approach), yeast-one-hybrid experiments on the specificity-changes of mutant DNA sequences(2,3), and other high-throughput experiments available such as SELEX(4), STARR-seq (5) or Sharpr-MPRA(6) (a modification of the MPRA(7,8) protocol that was developed to unveil at genome scale the effect of SNVs). Additional training and testing sets will be extracted from the datasets MANTA2 (9) and SNP2TFBS (10), composed by binding-site predictions in the human genome with the potential impact on TF binding for all possible SNVs. The impact of SNVs in MANTA2 is assessed by means of PWM scores computed on the alternate alleles. Similarly, we will use the theoretical PWMs plus all the energy profiles obtained by scanning the DNA sequence with the collection of TF-DNA structural models. Furthermore, the approach will be iteratively improved to be applicable on direct TF-DNA interactions in the human genome, using ChiP-seq data from recently expanded ENCODE encyclopaedia (11), updated versions of FactorBook (12), the dataset of UniBind (13) and datasets of altered cis-regulatory elements (14) and the location of DNase I hypersensitive regions of the genome with human genetic variation within transcription factor footprints(15,16)

References:

1. Bailey, T.L., Johnson, J., Grant, C.E. and Noble, W.S. (2015) The MEME Suite. Nucleic Acids Res, 43, W39-49. 2. Fuxman Bass, J.I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N., Yi, S., Hill, D.E., Vidal, M. and Walhout, A.J. (2015) Human gene-centered transcription factor networks for enhancers and disease variants. Cell, 161, 661-673. 3. Fuxman Bass, J.I., Pons, C., Kozlowski, L., Reece-Hoyes, J.S., Shrestha, S., Holdorf, A.D., Mori, A., Myers, C.L. and Walhout, A.J. (2016) A gene-centered C. elegans protein-DNA interaction network provides a framework for functional predictions. Mol Syst Biol, 12, 884. 4. Jolma, A., Yan, J., Whitington, T., Toivonen, J., Nitta, K.R., Rastas, P., Morgunova, E., Enge, M., Taipale, M., Wei, G. et al. (2013) DNA-binding specificities of human transcription factors. Cell, 152, 327-339. 5. Vockley, C.M., Guo, C., Majoros, W.H., Nodzenski, M., Scholtens, D.M., Hayes, M.G., Lowe, W.L., Jr. and Reddy, T.E. (2015) Massively parallel quantification of the regulatory effects of noncoding genetic variation in a human cohort. Genome Res, 25, 1206-1214. 6. Ernst, J., Melnikov, A., Zhang, X., Wang, L., Rogov, P., Mikkelsen, T.S. and Kellis, M. (2016) Genome-scale high-resolution mapping of activating and repressive nucleotides in regulatory regions. Nat Biotechnol, 34, 1180-1190. 7. Patwardhan, R.P., Lee, C., Litvin, O., Young, D.L., Pe'er, D. and Shendure, J. (2009) High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. Nat Biotechnol, 27, 1173-1175. 8. Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Jr., Kinney, J.B. et al. (2012) Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. Nat Biotechnol, 30, 271-277. 9. Fornes, O., Gheorghe, M., Richmond, P.A., Arenillas, D.J., Wasserman, W.W. and Mathelier, A. (2018) MANTA2, update of the Mongo database for the analysis of transcription factor binding site alterations. Sci Data, 5, 180141. 10. Kumar, S., Ambrosini, G. and Bucher, P. (2017) SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. Nucleic Acids Res, 45, D139-D144. 11. Consortium, E.P., Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A. et al. (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature, 583, 699-710. 12. Wang, J., Zhuang, J., Iyer, S., Lin, X.Y., Greven, M.C., Kim, B.H., Moore, J., Pierce, B.G., Dong, X., Virgil, D. et al. (2013) Factorbook.org: a Wiki-based database for transcription factor-binding data generated by the ENCODE consortium. Nucleic Acids Res, 41, D171-176. 13. Gheorghe, M., Sandve, G.K., Khan, A., Cheneby, J., Ballester, B. and Mathelier, A. (2019) A map of direct TF-DNA interactions in the human genome. Nucleic Acids Res, 47, e21. 14. Mathelier, A., Shi, W. and Wasserman, W.W. (2015) Identification of altered cis-regulatory elements in human disease. Trends Genet, 31, 67-76. 15. Vierstra, J., Lazar, J., Sandstrom, R., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Haugen, E. et al. (2020) Global reference mapping of human transcription factor footprints. Nature, 583, 729-736. 16. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D., Diegel, M., Dunn, D., Neri, F., Teodosiadis, A. et al. (2020) Index and biological spectrum of human DNase I hypersensitive sites. Nature, 584, 244-251. 17. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F. et al. (2017) Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science, 356. 18. Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E. and Tung, J. (2018) Genome-wide quantification of the effects of DNA methylation on human gene regulation. Elife, 7. 19. Meseguer, A., Årman, F., Fornes, O., Molina-Fernández, R., Bonet, J., Fernandez-Fuentes, N. and Oliva, B. (2020) On the prediction of DNA-binding preferences of C2H2-ZF domains using structural models: application on human CTCF. NAR Genomics and Bioinformatics, 2.

Expected skills::

Python programming

Possibility of funding::

No

Possible continuity with PhD: :

To be discussed

Universitat Pompeu Fabra Barcelona

Master in Bioinformatics for Health Sciences

Master project 2021-2022

Personal Information

Supervisor	Jana Selent
Email	jana.selent@upf.edu
Institution	IMIM-UPF
Website	www.jana-selent.org
Group	GPCR Drug Discovery Lab

Project

Structural bioinformatics

Project Title:

Detection of novel druggable binding sites at G protein-coupled receptors

Keywords:

G protein-coupled receptors, molecular dynamics, data analysis, drug design

Summary:

G protein-coupled receptors (GPCRs) are the most abundant class of receptors in the human organism. They are present in almost every type of cell, and govern almost every process in the human body (i.e. cognitive and inflammatory processes or control of the cardiovascular system). Owing to their ubiquity, they are targets of more than 30% of current drugs, and every day new GPCRs are revealed to be pharmacological targets for existing diseases. Molecular dynamics (MD) is a sophisticated technique that enables to simulate protein behaviour in a physiological environment. In this project the Master student will develop a simulation-based pipeline that allows detecting druggable binding sites including cryptic pockets ("transient binding pockets"). This pipeline involves (i) the setup of simulation systems including small chemical fragments to probe the entire protein surface for druggable binding sites, (ii) running production runs, (iii) automated detection of binding sites and (iii) intuitive visualization. The pipeline will be implemented into our GPCRmd server and detected sites will be exploited for the discovery of new molecular GPCR modulators. We expect that the results of the analysis will be published in a high impact journal, and the expertise acquired by the student will make her/him a valuable asset for pharma companies in future. We are looking for a highly motivated and skilled student with exceptional academic records that allows pursuing a PhD afterwards.

References:

Rodríguez-Espigares & Torrens-Fontanals et al. GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods 2020, DOI: 10.1038/s41592-020-0884-y

Expected skills::

Experience in structural biology, programming in python/bash, molecular dynamics engines (GROMACS, NAMD, etc.), analysis tools/packages (VMD, Chimera, MDtraj…) and high level of English, oral and written.

Possibility of funding::

To be discussed

Possible continuity with PhD: :

Yes