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