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