Università degli Studi di Verona

DIPARTIMENTO DI INFORMATICA
Ph.D. in Computer Science

First Year Report

Predicting genetic variants effect on genomic Regulatory Elements

Student:
Manuel Tognon
Matricola VR456869

Supervisor:
Prof. Rosalba Giugno

Cosupervisor:
Prof. Luca Pinello

Anno Accademico 2020–2021

Contents

4

List of Figures

List of Tables

Introduction

Transcription Factors (TFs) are fundamental regulatory proteins playing a key role in regulating the transcriptional state, differentiation and developmental patterns of cells (Lambert *et al.*, 2018; Reimold *et al.*, 2001; Whyte *et al.*, 2013). By binding short DNA sequences (7-20 nucleotides (Stewart *et al.*, 2012)) called transcription factor binding sites (TFBS) they finely regulate gene expression in a cell-specific manner. TFBS are located within gene promoters (Whitfield *et al.*, 2012) or in distal regulatory elements, such as enhancers or silencers (Gotea *et al.*, 2010; Lemon and Tjian, 2000; Nolis *et al.*, 2009). TFs bind DNA in a sequence specific manner, recognizing similar but not identical sequences differing in few nucleotides. Often TFBS of a given TF show recurring patterns, which are referred to as *motifs*. TFBS discovery or *motif discovery* is one of the most studied and challenging problems in genomics and computational genomics (Pavesi *et al.*, 2004; D'haeseleer, 2006; Zambelli *et al.*, 2013). TFBS motif discovery can be defined as the problem of finding short similar nucleotide patterns, shared by all or large fractions of sequences bound by the same TF, building the motif. TF motifs can be described and predicted by several models, such as Position Weight Matrices (PWMs) (Stormo, 2000), Markov models (MMs) (Durbin *et al.*, 1998), or Deep Neural Networks (DNNs) (Talukder *et al.*, 2021). During the last two decades, have been introduced several experimental methods to identify and characterize TFBS *in vitro* and *in vivo* (Jolma and Taipale, 2011), such as protein binding microarray (PBM) (Berger *et al.*, 2006; Berger and Bulyk, 2009), HT-SELEX (Jolma *et al.*, 2010), ChIP on Chip (Pillai and Chellappan, 2015; Collas and Dahl, 2008), or ChIP-seq (Johnson *et al.*, 2007; Mardis, 2007). These methods provide two major advantages: (i) they do not require any prior knowledge on binding site sequence, and (ii) they produce huge datasets of thousands of sequences bound by the studied TF. However, the actual binding sites remain to be computationally discovered. Several studies showed that genetic variants can significantly impact TF-DNA binding affinity (De Gobbi *et al.*, 2006; Weinhold *et al.*, 2014; Guo *et al.*, 2018). Genome-wide association studies (GWASs) uncovered thousands of genetic variants (SNPs) associated with complex human traits. The majority of identified SNPs are in non coding regions, often corresponding to functional regulatory elements, such as enhancers (Maurano *et al.*, 2012). This suggests that gene misregulation may be mediated by SNPs modulating TF-DNA binding interactions. In fact, these variants may perturb TF-DNA binding specificity, ultimately changing downstream gene expression (Deplancke *et al.*, 2016). Importantly, mutations altering TFBS can occur in haplotypes conserved within a population of individuals (Kasowski *et al.*, 2010), producing population specific TFBS motifs. Similarly, cell-type specific genetic variation can produce different motifs for the same TF. Therefore, developing new computational methods enabling haplotype- and variant-aware motif discovery is fundamental to describe genetic variation impact on TFBS at population level. Moreover, it is important that such models are easily interpretable by humans.

References

Berger, M. F. and Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the dna-binding specificities of transcription factors. *Nature protocols*, **4**(3), 393–411.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006). Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, **24**(11), 1429–1435.

Collas, P. and Dahl, J. A. (2008). Chop it, chip it, check it: the current status of chromatin immunoprecipitation. *Frontiers in Bioscience-Landmark*, **13**(3), 929–943.

De Gobbi, M., Viprakasit, V., Hughes, J. R., Fisher, C., Buckle, V. J., Ayyub, H., Gibbons, R. J., Vernimmen, D., Yoshinaga, Y., De Jong, P., *et al.* (2006). A regulatory snp causes a human genetic disease by creating a new transcriptional promoter. *Science*, **312**(5777), 1215–1217.

Deplancke, B., Alpern, D., and Gardeux, V. (2016). The genetics of transcription factor dna binding variation. *Cell*, **166**(3), 538–554.

D'haeseleer, P. (2006). How does dna sequence motif discovery work? *Nature biotechnology*, **24**(8), 959–961.

Durbin, R., Eddy, S. R., Krogh, A., and Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Gotea, V., Visel, A., Westlund, J. M., Nobrega, M. A., Pennacchio, L. A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. *Genome research*, **20**(5), 565–577.

Guo, Y. A., Chang, M. M., Huang, W., Ooi, W. F., Xing, M., Tan, P., and Skanderup, A. J. (2018). Mutation hotspots at ctcf binding sites coupled to chromosomal instability in gastrointestinal cancers. *Nature communications*, **9**(1), 1–14.

Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-dna interactions. *Science*, **316**(5830), 1497–1502.

Jolma, A. and Taipale, J. (2011). Methods for analysis of transcription factor dna-binding specificity in vitro. *A Handbook of Transcription Factors*, pages 155–173.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J. M., Yan, J., Sillanpää, M. J., *et al.* (2010). Multiplexed massively parallel selex for characterization of human transcription factor binding specificities. *Genome research*, **20**(6), 861–873.

Kasowski, M., Grubert, F., Heffelfinger, C., Hariharan, M., Asabere, A., Waszak, S. M., Habegger, L., Rozowsky, J., Shi, M., Urban, A. E., *et al.* (2010). Variation in transcription factor binding among humans. *science*, **328**(5975), 232–235.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, **172**(4), 650–665.

Lemon, B. and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. *Genes & development*, **14**(20), 2551–2569.

Mardis, E. R. (2007). Chip-seq: welcome to the new frontier. *Nature methods*, **4**(8), 613–614.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., *et al.* (2012). Systematic localization of common disease-associated variation in regulatory dna. *Science*, **337**(6099), 1190–1195.

Nolis, I. K., McKay, D. J., Mantouvalou, E., Lomvardas, S., Merika, M., and Thanos, D. (2009). Transcription factors mediate long-range enhancer–promoter interactions. *Proceedings of the National Academy of Sciences*, **106**(48), 20222–20227.

Pavesi, G., Mauri, G., and Pesole, G. (2004). In silico representation and discovery of transcription factor binding sites. *Briefings in bioinformatics*, **5**(3), 217–236.

Pillai, S. and Chellappan, S. P. (2015). Chip on chip and chip-seq assays: genome-wide analysis of transcription factor binding and histone modifications. In *Chromatin Protocols*, pages 447–472. Springer.

Reimold, A. M., Iwakoshi, N. N., Manis, J., Vallabhajosyula, P., Szomolanyi-Tsuda, E., Gravallese, E. M., Friend, D., Grusby, M. J., Alt, F., and Glimcher, L. H. (2001). Plasma cell differentiation requires the transcription factor xbp-1. *Nature*, **412**(6844), 300–307.

Stewart, A. J., Hannenhalli, S., and Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, **192**(3), 973–985.

Stormo, G. D. (2000). Dna binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Talukder, A., Barham, C., Li, X., and Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Briefings in Bioinformatics*, **22**(3), bbaa177.

Weinhold, N., Jacobsen, A., Schultz, N., Sander, C., and Lee, W. (2014). Genome-wide analysis of noncoding regulatory mutations in cancer. *Nature genetics*, **46**(11), 1160–1165.

Whitfield, T. W., Wang, J., Collins, P. J., Partridge, E. C., Aldred, S. F., Trinklein, N. D., Myers, R. M., and Weng, Z. (2012). Functional analysis of transcription factor binding sites in human promoters. *Genome biology*, **13**(9), 1–16.

Whyte, W. A., Orlando, D. A., Hnisz, D., Abraham, B. J., Lin, C. Y., Kagey, M. H., Rahl, P. B., Lee, T. I., and Young, R. A. (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, **153**(2), 307–319.

Zambelli, F., Pesole, G., and Pavesi, G. (2013). Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in bioinformatics*, **14**(2), 225–237.