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

Introduction **9**

Background on Motif Discovery **11**

2.1 Motif Discovery methods 11

List of Figures

2.1 Schematic representation of motif discovery workflow. 12

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.*, 2004a; 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.

Background on Motif Discovery

TFBS motif discovery can be defined as follows. Given a set of nucleotide sequences S produced by experimental assays (**Fig.2.1**) sharing a common biological function, find one or more motifs, built with one or more sets of short similar patterns, appearing in a large fraction of S. Therefore, it is fundamental to allow for experimental errors and potential false positive in S. To assess motifs statistical significance, they should not appear with similar frequency in sets of background DNA sequences B. Moreover, patterns building the motif should not have the same degree of similarity found in $b_i \in B$. Motifs can be encoded and represented with different models M. Motif model choice and construction is fundamental. In fact, M is often used to predict new potential occurrences of TFBS motifs in DNA sequences, not used during model training procedure, and potentially assess the effects of genetic variants on TF-DNA binding affinity. The various methods introduced to solve this challenge mainly differ on three points: (i) the method employed to derive the motif, (ii) the model chosen to represent the motif, and (iii) how statistical significance of motifs is assessed and which background model is used (Zambelli *et al.*, 2013).

2.1 Motif Discovery methods

Motif discovery methods can be classified in five classes, based on the algorithm employed to discover putative TFBS motifs: consensus sequences, alignment profiles, markov models (MM), support vector machines (SVM), and deep neural networks (**Fig.2.1 (B)**). Algorithms employing consensus sequence methods, collect all the approximate occurrences (with up to e mismatches) of each of 4^m nucleotide sequences of length m (m is the motif width, $\sim 10 - 20$ bp) in the input sequences, counting the number of matches. Therefore, the general idea is to enumerate all possible DNA oligos and count how many times they are observed in the input sequence set S, with respect to the background dataset B. The most enriched and significant motifs are reported as potential binding sites and are furtherly used to build the TFBS model. This basic approach was employed in early studies on TFBS characterization (Waterman *et al.*, 1984; Galas *et al.*, 1985), although shown to be very computationally demanding. The application of indexed data structures, such as suffix trees (Välimäki *et al.*, 2007) made the approach feasible in real-world studies. Moreover, employing suffix trees the algorithm complexity becomes exponential in the number of e mismatches allowed instead on motif width m (Zambelli *et al.*, 2013). Weeder (Pavesi *et al.*, 2001, 2004b) and SMILE (Marsan and Sagot, 2000) algorithms extended the approach by performing an exhaustive matching with no restrictions on substitution positions. While SMILE compares motif occurrences frequency with those expected in B, Weeder checks motif occurrences in S against oligos expected frequencies in all promoter regions of the same organism of input sequence set. The selected oligos are furtherly encoded in PWMs, to describe the putative motifs.

Alignment profiles provide a more powerful and flexible description of motif binding preferences. Therefore, they have been the basic idea of choice of many motif discovery methods. The general idea is to construct an alignment profile with oligos selected from S and score the resulting profile according to nucleotide conservation and with a suitable measure of significance. The problem can be formalized as a combinatorial optimization, searching for the best possible combination of fixed-length motif patterns. Therefore, the goal is to find the highest scoring profile exploring all the possible alignments of fixed-length sequences from S. The most naive solution would be to exhaustively enumerate all possible alignments, but it would be too computationally demanding. To solve this challenge, alignment profile-based algorithms employ different types of heuristics and combinatorial optimizations, such as Expectation-Maximization (EM) (Bailey and Elkan, 1995a), greedy (Hertz and Stormo, 1999), stochastic (Lawrence *et al.*, 1993), and genetic algorithms (Lee *et al.*, 2018). MEME algorithm (Bailey *et al.*, 1994; Bailey and Elkan, 1995b; Bailey *et al.*, 2006) explores the solution space employing an EM optimization strategy. Given a starting profile, MEME iteratively substitutes some sequences of the profile with others likely to produce better solutions. Therefore, aligned sequences are more likely to fit the profile than

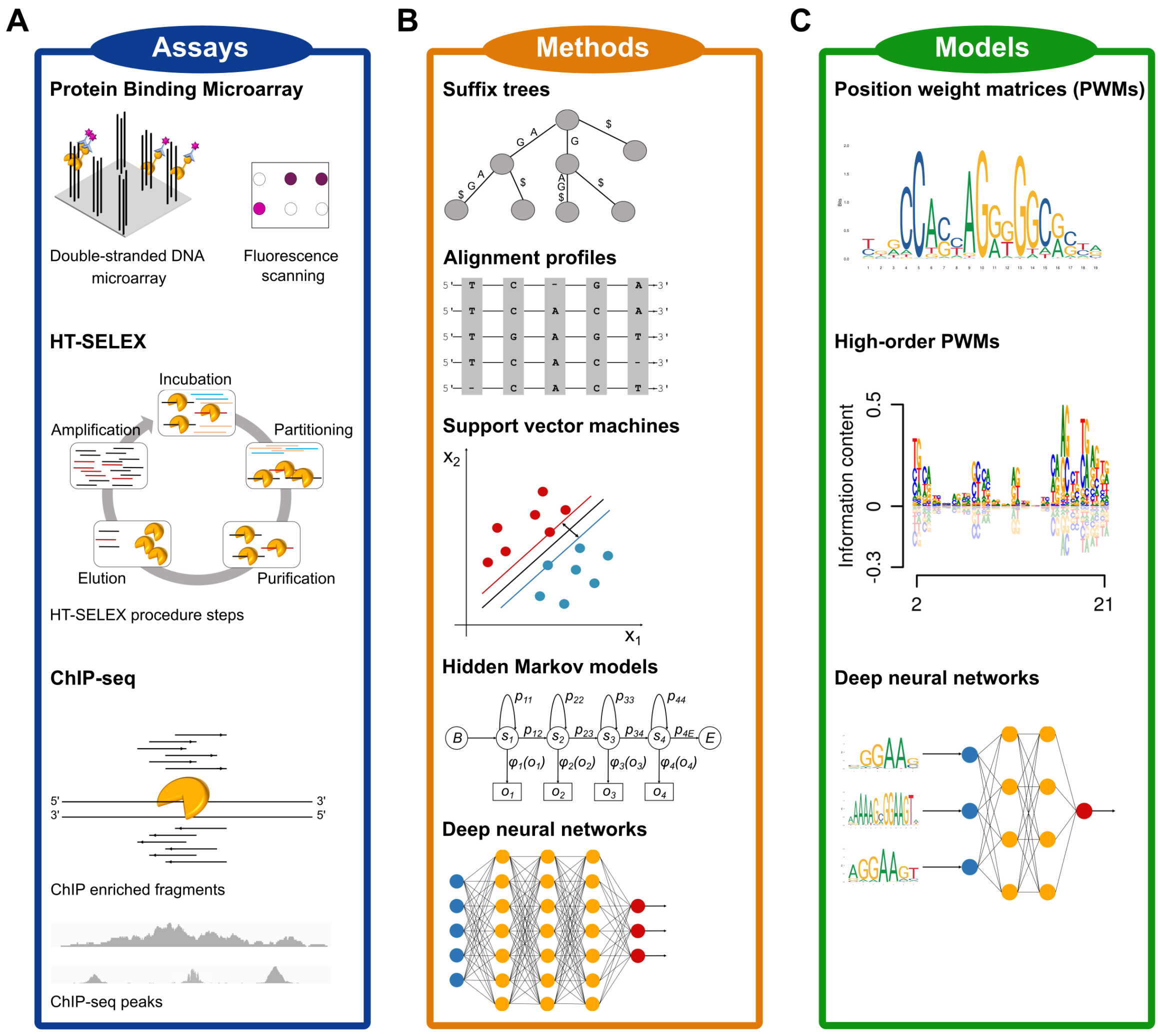

Figure 2.1. Schematic representation of motif discovery workflow. **(A)** Experimental assays identify and characterize nucleotide sequences containing binding sites of the studied TF. The actual binding site yet remains to be discovered and characterized. Sequences identified by experimental assays constitute the input of motif discovery algorithms. **(B)** Motif discovery methods can be divided into five major classes based on the employed algorithm: consensus sequence (suffix trees), alignment profiles, support vector machines, markov models, deep neural networks. **(C)** TFBS motifs are described and summarized in models built using patterns identified by algorithms, such as PWM, DWM or neural networks.

the remaining oligos, which should fit a background model better than the profile. However, MEME EM local search strategy can reach premature convergence in local maxima. Stochastic optimization strategies, such as Gibbs sampling, attempt to avoid this important limitation [42]. While MEME build the staring alignment profile with all fixed length sequences from S, stochastic algorithms build the profile choosing randomly an oligo from each $s \in S$. Then, the sequence from each $s_i \in S$ is removed from the profile and a likelihood score is computed for each oligo in s_i describing how well it fits the profile, rather than a background model. The removed oligo is replaced by another sequence from s_i with probability proportional to the computed likelihood score. The procedure is iteratively repeated until convergence is reached, or after a fixed number of iterations. It is important to notice that both local and stochastic search strategies assume the one binding site appears in each input sequence. Motif sampler [43] extended the stochastic search strategy allowing for multiple or no occurrences of a motif in input sequences. Modifications of the basic Gibbs sampling procedure were introduced in AlignACE [44] and ANN-Spec, which pairs sampling procedure with an artificial neural network [45]. GLAM algorithm [46] furtherly improved stochastic optimization, by allowing the comparison between motifs of different lengths through a simulated annealing strategy. The method was furtherly improved in GLAM2 [47] to consider gaps within motifs.

References

Bailey, T. L. and Elkan, C. (1995a). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine learning*, **21**(1), 51–80.

Bailey, T. L. and Elkan, C. (1995b). The value of prior knowledge in discovering motifs with meme. In *Ismb*, volume 3, pages 21–29.

Bailey, T. L., Elkan, C., *et al.* (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers.

Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). Meme: discovering and analyzing dna and protein sequence motifs. *Nucleic Acids Research*, **34**(suppl_2), W369–W373.

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.

Galas, D. J., Eggert, M., and Waterman, M. S. (1985). Rigorous pattern-recognition methods for dna sequences: Analysis of promoter sequences from escherichia coli. *Journal of molecular biology*, **186**(1), 117–128.

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.

Hertz, G. Z. and Stormo, G. D. (1999). Identifying dna and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, **15**(7), 563–577.

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.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a gibbs sampling strategy for multiple alignment. *science*, **262**(5131), 208–214.

Lee, N. K., Li, X., and Wang, D. (2018). A comprehensive survey on genetic algorithms for dna motif prediction. *Information Sciences*, **466**, 25–43.

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.

Marsan, L. and Sagot, M.-F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of computational biology*, **7**(3-4), 345–362.

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. (2001). An algorithm for finding signals of unknown length in dna sequences. *Bioinformatics*, **17**(suppl_1), S207–S214.

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

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004b). Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic acids research*, **32**(suppl_2), W199–W203.

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.

Välimäki, N., Gerlach, W., Dixit, K., and Mäkinen, V. (2007). Compressed suffix tree—a basis for genome-scale sequence analysis. *Bioinformatics*, **23**(5), 629–630.

Waterman, M., Arratia, R., and Galas, D. (1984). Pattern recognition in several sequences: consensus and alignment. *Bulletin of mathematical biology*, **46**(4), 515–527.

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.