Complexity of transcription factor binding motifs

Jan Grau[1], and Jens Keilwagen[2]

[1]Institute of Computer Science, Martin Luther University Halle–Wittenberg, Germany

[2]Julius Kühn-Institut (JKI) - Federal Research Centre for Cultivated Plants, Quedlinburg, Germany

1 Introduction

Transcriptional regulation mediated by transcription factors (TFs) binding to genomic DNA is one of the fundamental regulatory steps of gene expression. Over the last years, the importance of dependencies between different positions of transcription factor binding sites (TFBSs) has been debated controversially [1, 13, 9]. Several publications argue that TF-DNA binding energies can often be captured by simple weight matrices [13, 11], whereas others find that considering dependencies increases the performance of TFBS predictions [8, 7, 6, 4].

Here, we aim at providing new insights into the importance of dependencies in transcription factor binding sites and investigate the diverse sources of such dependencies on *in-vitro* genomic context protein binding microarray (gcPBM) data [8] and *in-vivo* ChIP-seq data from ENCODE [10]. For this purpose, we propose a new class of probabilistic models that allow for learning dependencies between binding site positions discriminatively, which we call *sparse local inhomogeneous mixture* (Slim) models. For representing dependencies graphically, we develop a new visualization technique, which we call *dependency logos*.

2 Sparse local inhomogeneous Markov models

Determining a probabilistic model requires the selection of features and the estimation of corresponding model parameters. Typically, feature selection is performed in discrete space (features are selected or not), while parameter estimation is performed in continuous space. For parameter estimation, discriminative learning principles have been proven superior over generative ones in many areas including motif discovery [2, 5, 4], but typically demand for time-consuming numerical optimization, which makes them intractable for traditional feature selection that requires a new optimization for each (promising) feature subset.

To overcome this situation, we propose Slim models that use the alternative concept of soft feature selection. More specifically, the probability of a nucleotide at a certain position of a binding site may depend on any nucleotide observed at a preceding position. Since it is unknown beforehand, which of these putative dependencies are important, the Slim model handles this information as a hidden variable resulting in a local mixture model. During the learning process, the parameters of this mixture model are adapted, such that a single position or a small subset of preceding positions obtains a large weight, whereas the others are down-weighted, yielding a soft feature selection.

3 Dependency logos

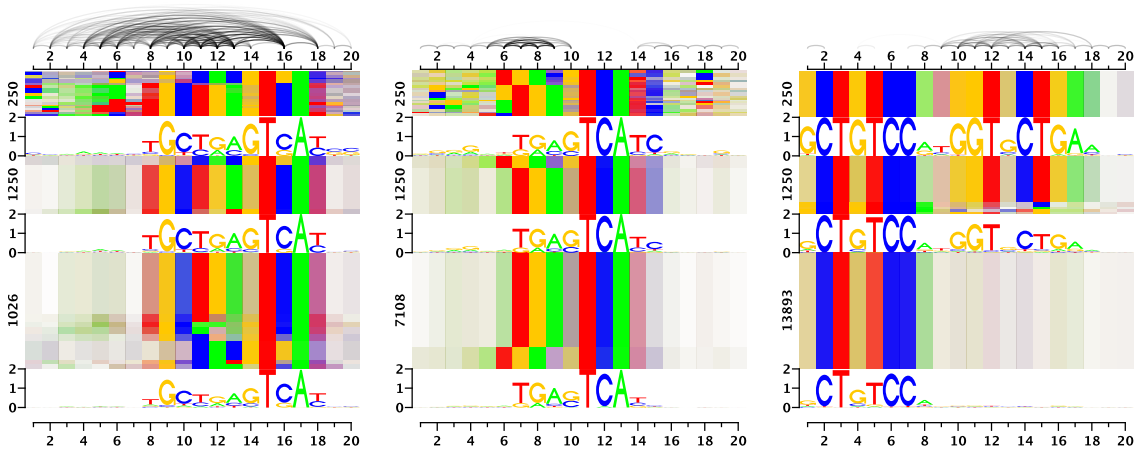
We present dependency logos as a new way of visualizing dependency structures within binding sites. In contrast to sequence logos, dependency logos make dependencies between binding site positions visually perceptible. In contrast to previous approaches, dependency logos are model-free and only require a set of aligned sequences, e.g., predicted binding sites, and, optionally, associated weights as input.

Dependency logos make dependencies between different motif positions visually perceptible by three key ideas. First, dependency logos are directly based on binding sites instead of abstract binding motifs, e.g., mononucleotide distributions of PWM models. Second, we cluster binding sites by their nucleotides at those positions showing the strongest dependencies to other positions. If, for instance, position i shows the strongest dependencies to other positions and, of those, the dependency between position j and i is the strongest, we create at most 16 clusters according to the combinations of the two nucleotides present at positions j and i. This procedure may be repeated recursively for each of the clusters (e.g., those sequences with a TC at position j and i). Third, we visualize each cluster as one row of colored boxes using the familiar colors of sequence logos and with height proportional to cluster size. If more than one nucleotide is present at a certain binding site position in a cluster, we mix the colors representing those nucleotides and set their saturation based on information content in analogy to the height of stacks in sequence logos.

4 Results

We demonstrate that Slim models in combination with a discriminative learning principle yield an overall improved performance compared to state of the art tools and compared to other probabilistic models including position weight matrix models on gcPBM and 63 ChIP-seq data for human transcription factors. Scrutinizing the results of the individual data sets, we find several cases where a PWM model neglecting dependencies between binding site positions already yields a decent prediction performance. However, for a considerable fraction of data sets, the improvement gained by models capturing dependencies between adjacent and non-adjacent positions is substantial.

Subsequently, we focus on ChIP-seq data sets for those transcription factors with the greatest improvements in prediction performance using Slim models and further investigate their dependency structures using dependency logos. In Figure 1, we show three examples of dependency logos based on predictions of Slim models. For Nfe2, we observe heterogeneities caused by two different, mixed



(a) Nfe2 (b) c-Jun (c) Nrsf

Figure 1: Dependency logos of binding sites predicted by the Slim model for ChIP-seq data sets.

motifs, where the first is an E-box-like (CACGTG) motif and the second is the expected Nfe2 motif with consensus TGCTGAGTCAY. For c-Jun, we find a flexible spacer between the two half sites with consensus TGA and TCA that has also been reported by Badis *et al.* [1] for Jundm2 in mouse using PBM data and by Mathelier and Wasserman [7] using TFFMs on ChIP-seq data for human Jund. For Nrsf, we find that only the top-scoring binding sites cover the complete Nrsf motif, whereas the majority of sequences under the ChIP-seq peaks (68%) contain only the left half site (CTGTCC). While a dependency of nucleotide conservation on ChIP enrichment of the Nrsf motif has been reported before [3], the clear distinction between two modes of Nrsf binding discovered using the Slim model is novel and might be related to the diverse complexes of Nrsf with other factors [12].

In summary, we find that binding landscapes of transcription factors are highly complex and diverse, including secondary or multiple motifs, partial motifs, flexible binding modes, or dependencies between neighboring and non-neighboring positions. Some of these cases could also be handled by specialized models based on *a-priori* expert knowledge, e.g., spaced PWM models for c-Jun or hidden Markov model-like approaches for Nrsf. The strength of the proposed Slim models is their flexibility to adjust to all these dependency structures without requiring *a-priori* knowledge of dependency structures, while dependency logos allow for dissecting dependency structures *a-posteriori* by visual inspection.

References

[1] Gwenael Badis et al. Diversity and complexity in DNA recognition by transcription factors. *Science*, 324(5935):1720–1723, 2009.

[2] Timothy L. Bailey. DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, 27(12):1653–1659, 2011.

[3] Alexander W. Bruce et al. Functional diversity for REST (NRSF) is defined by in vivo binding affinity hierarchies at the DNA sequence level. *Genome Research*, 19(6):994–1005, 2009.

[4] Jan Grau et al. A general approach for discriminative de novo motif discovery from high-throughput data. *Nucleic Acids Research*, 41(21):e197, 2013.

[5] Peter Huggins et al. DECOD: fast and accurate discriminative DNA motif finding. *Bioinformatics*, 27(17):2361–2367, 2011.

[6] Ivan Kulakovskiy et al. From binding motifs in ChIP-seq data to improved models of transcription factor binding sites. *Journal of Bioinformatics and Computational Biology*, 11(01):1340004, 2013.

[7] Anthony Mathelier and Wyeth W. Wasserman. The next generation of transcription factor binding site prediction. *PLoS Comput Biol*, 9(9):e1003214, 09 2013.

[8] Fantine Mordelet et al. Stability selection for regression-based models of transcription factor–DNA binding specificity. *Bioinformatics*, 29(13):i117–i125, 2013.

[9] Quaid Morris et al. Jury remains out on simple models of transcription factor specificity. *Nat Biotech*, 29(6):483–484, 06 2011.

[10] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 09 2012.

[11] Matthew T. Weirauch et al. Evaluation of methods for modeling transcription factor sequence specificity. *Nature biotechnology*, 31(2):126–134, February 2013.

[12] Hong-Bing Yu et al. Coassembly of REST and its cofactors at sites of gene repression in embryonic stem cells. *Genome Research*, 21(8):1284–1293, 2011.

[13] Yue Zhao and Gary D. Stormo. Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature biotechnology*, 29(6):480–483, June 2011.