A PROBABILISTIC MODEL FOR COMPETITIVE BINDING OF TRANSCRIPTION FACTORS

Kirsti Laurila[1] *and Harri Lähdesmäki*[1,2]

[1]Department of Signal Processing, Tampere University of Technology,
P.O. Box 527, FI-33101 Tampere, Finland
[2]Department of Information and Computer Science, Helsinki University of Technology,
P.O. Box 5400, FI-02015 TKK, Finland
kirsti.laurila@tut.fi, harri.lahdesmaki@tut.fi

ABSTRACT

One of the most important regulation steps of gene expression is transcriptional regulation, which is to a large extent controlled by transcription factors binding to DNA. Many models of transcription factor binding have been proposed but most of them model binding of a single transcription factor at a time. Existing prediction methods for multiple TFs base mainly on searching for clustered binding sites or cis-regulatory modules. We have developed a probabilistic model that predicts simultaneously binding of several transcription factors. Our method considers the transcription factor binding process as a competition between factors which is realistic from the biological point of view. Modeling results show a remarkable improvement compared to the cases where individual binding prediction results of separate TFs have been combined.

1. INTRODUCTION

Gene expression is regulated in several steps of which transcription regulation is one of the most important ones. Thus, understanding the transcription factor (TF) binding on DNA is of great importance as it is the main step in transcription control. Existing methods such as chromatin immunoprecipitation-chip (ChIP-chip) or -sequencing (ChIP-seq) techniques allow us to study TF binding experimentally. Nevertheless, these methods allow to study binding of only one TF at a time in certain conditions. Further, experimental methods are time consuming and cannot be used to screen all TFs because specific antibodies do not exist. As a consequence, predicting transcription factor binding sites (TFBSs) computationally is important in order to understand the whole transcription and gene expression regulation processes.

Many TFBS prediction methods exists but most of them predict the binding sites of only a single TF at a time. This kind of prediction methods perform usually with a good sensitivity but they lack specificity and, thus, give a large set of false TFBS predictions. This makes it hard to determine the real binding sites, especially because the real TFBS can be strong or weak ones. Further, transcription of a single gene is controlled by several TFs and thus one needs to perform individual prediction for each of these regulating TFs — this complicates the problem further and increases the number of false positive predictions. However, predicting individually the binding of each TF is not realistic as it does not consider the interactions between different proteins and other molecules present in the cell. These other factors can interrupt already bound TFs or they can prevent binding of some TFs even though it has a strong affinity to its binding site. Thus predicting the binding of all different TFs at the same time allows considering DNA binding as a competition between the factors and mimics the biology in the cell better than combining the predictions of individual predictions.

In addition to the methods for predicting a single TF at a time, some models predicting TFBS of multiple TFs also exist. These methods can usually be divided into two categories [1]. As two different methods presented in [2, 3], the first category consists of methods that search for TFBS that are located close to each other. The basic principle behind this kind of methods is the fact that proximal binding sites makes the interactions between TFs possible which is can be essential for the transcription process. The second category of methods search for so called cis-regulatory modules which are clusters of binding sites of cooperative TFs. This kind of methods are presented for example in [4, 5].

In this paper, we propose a new method that predicts TFBS of several TFs simultaneously. This method builds on the standard probabilistic sequence specificity models, combines them into an integrated model and makes Bayesian inference for binding sites [6]. Modeling results show remarkable improvement compared to the cases where the individual prediction results of separate TF binding have been combined.

2. METHODS

We formulate a probabilistic model for competitive TF binding prediction. The goal is to develop a realistic model that takes into account simultaneous binding of several TFs to the same DNA sequence and hence explicitly model competition of binding sites by several TFs. Instead of using deterministic binding sites, our approach makes use of the known fact that almost all TFs can bind to any DNA

sequence stretch, the strength of binding being determined by the sequence affinity of each TF. Thus, the method automatically models both weak and strong binding sites, both of which are known to be important for transcription regulation. Our modeling framework is probabilistic and all quantities and phenomena, including the TF binding itself, can be answered in terms of probabilities which naturally represents our belief in TFBSs.

First, we model the binding of m TFs to a whole ℓ-length promoter sequence $S = (S_1, \ldots, S_\ell)$, where $S_i \in \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$. Binding affinity of each TF i is represented by a position specific frequency matrix (PSFM) Θ_i and non-binding sites are represented by the standard d-order Markovian background model ϕ (we use $d = 3$). PSFMs for m TFs are collectively denoted by $\Theta = (\Theta_1, \ldots, \Theta_m)$. To model binding of multiple TFs simultaneous, let $A = \{a_1, \ldots, a_c\}$ denote the starting positions of c non-overlapping binding sites on S, and vector $\pi = (\pi_1, \ldots, \pi_c)$ specifies the numerical labels of bound TFs (i.e., $\pi_i \in \{1, \ldots, m\}$). Using these definitions alone, it is straightforward to compute the conditional probability of a sequence $P(S|A, \pi, \Theta, \phi)$.

Instead of the standard frequentist computation, we implement a Bayesian alternative that associates the PSFMs and the Markovian background model with Dirichlet priors, integrates out the model parameters and computes the posterior binding probability of the given TFBSs $P(A, \pi|S) \propto P(S|A, \pi)P(A, \pi)$. Instead of searching for the maximum a posteriori (MAP) binding site configuration, we compute Bayesian posterior probabilities for all TFBS locations, resulting in the full posterior $P(A, \pi|S)$ over A and π. We sum the probabilities of relevant locations as

$$P(\text{“TF } k \text{ binds the promoter sequence } S\text{”}) = \sum_{(A,\pi)\in\mathcal{A}_k} P(A, \pi|S), \text{ where}$$
$$\mathcal{A}_k = \{(A, \pi) : \exists i : \pi_i = k\}.$$

This gives a straightforward way of computing posterior binding probabilities for any TF-promoter pair. Note that although binding sites are non-overlapping for a fixed (A, π), in the above equation we sum over all possible pairs (A, π). In other words, for a promoter S, we consider all possible numbers of binding sites in any possible configuration. This implicitly takes into account the possibility that two or more TFs can bind the same genomic location but binding cannot happen simultaneously (i.e., competition for the binding sites).

Direct computation of the Bayesian probability for all TFBS locations is intractable. To this end, we devise our model with a flexible Markov chain Monte Carlo (MCMC) estimation method. This model allows modeling competitive binding of any number of TFs to DNA. We use a similar Metropolis Hastings (MH) algorithm as in [6] but adapted to the problem of multiple TFs. The MH algorithm iteratively proposes to either add a new TFBS or delete an existing TFBS, and the proposed moves are accepted with the probability that satisfies the detailed balance condition.

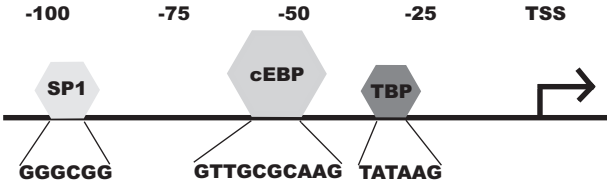

Figure 1. Known binding sites for mouse leptin (U36238) promoter. TSS=transcription starting site.

The test set used in the study was obtained by filtering the test set used in [6]. This test set was collected from ABS [7] and ORegAnno[8] databases. To get a test set suitable for this study, we removed the sequences for which we had only one binding site or the PSFMs for the TFs were known only for one TF. After this we had 29 promoter sequences. Promoter sequences were cut 50 nucleotides before the first known TFBS and 50 nucleotides after the last one. The lengths of sequences varied between 110 and 1322 nucleotides.

We collected a set of non-redundant PSFMs from TRANSFAC [9] (Release 10.3) which were used to construct the Dirichlet priors for the TFs. For the 29 promoter sequences, $m = 27$ different TFs had binding sites and this TF set was used for every promoter. Note that, for a specific promoter, this set typically contains several TFs that do not bind the promoter. Markov model parameters were estimated from an additional set of 250 nucleotides long upstream non-coding sequences (both strands) [6].

3. RESULTS

We computed the TFBS predictions for our test set with our competitive model integrating all 27 TFs in one prediction and compared the results with the results where individual predictions for each TF [6] were combined. The results showed remarkable improvement as with combined individual predictions one got many TFs to bind to the same DNA sequence which is usually physically impossible. An example shown in Figure 1 illustrates a part of the mouse leptin promoter and its known TFBSs. Three different TFs are known to bind to this promoter, SP1 to region -100/-95 (relative to transcription starting site), cEBP to region -58/-28 and TBP to the region -33/-28 [10]. Other TFs also have high affinity to the cEBP binding site, namely MYB, SP1 and TEAD.

Binding predictions for the region -145/-1 of the leptin promoter are shown in Figure 2 which show the TFBS predictions for those TFs that are known to have binding sites on the promoter or are predicted to bind. With individual predictions, one predicts right the SP1 binding site but for cEBP binding site predictions suggest that in addition to cEBP, also MYB, SP1 and TEAD could indeed bind this site (see Figure 2a). Note well that combining individual predictions for several TFs generates a complicated end result where a number of TFs bind to the gene promoter, and the problem becomes increasingly severe with the increasing number of TFs. This problem is one of the

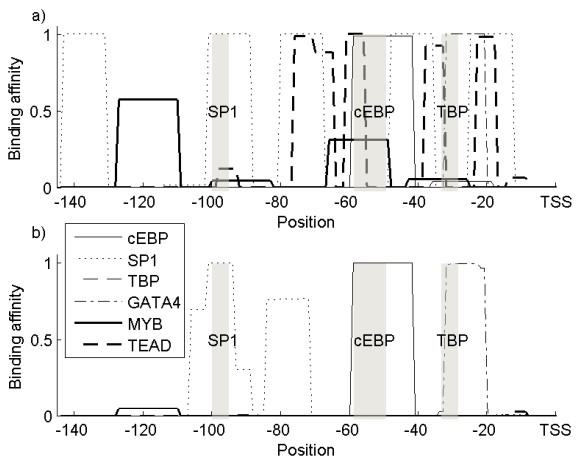

Figure 2. Predictions for mouse leptin promoter. Predictions are presented for those TFs that bind with affinity more than 0.1. Known binding sites are shaded. a) Individual predictions b) Competitive model.

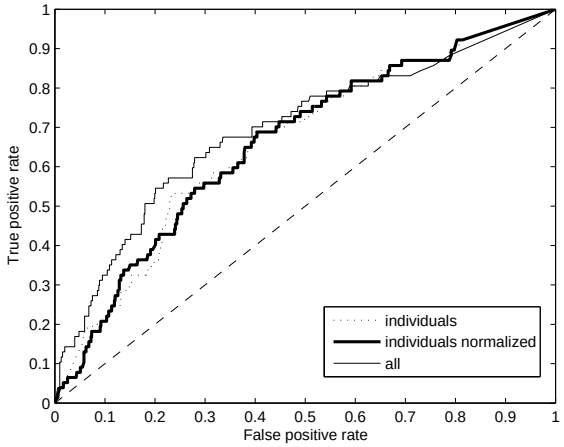

Figure 3. ROC curves for different methods. Only binding for promoter matters.

reasons why traditional TFBS prediction methods have an overwhelming number of false positives. When using the proposed model which explicitly models competitive binding, the binding of cEBP is predicted correctly (see Figure 2b). Perhaps more importantly, with competitive model one also gets only two false TFBS predictions (and few very weak binding sites) whilst with combined individual predictions there exists over ten false TFBS predictions and it is impossible to say which of them are correct. Either of the methods could not predict TBP's binding site which is most probably because of an incompatible binding specificity model. With other sequences the results were similar. Combined individual predictions showed the same binding site predictions than the multiTF but also a large number of other predictions that overlapped and contradicted.

We have also evaluated the comprehensive usefulness of our methods with receiver operating characteristic curves (ROC) and the area under the curve (AUC). The ROC curves (in Figure 3) show a large improvement in the performance of the competitive method relative to the combined individual predictions. Especially with small false positive rates (FPR) (which are the most preferable ones) the competitive method show to be far more discriminative from the random case than the individual predictions. The same result can be found from the AUC scores in Table 1. For FPR 0.1 the AUC of the competitive model is over 40% bigger than that for combined individual predictions. In Figure 3 and in Table 1 are also presented the values and plots for predictions that are obtained by normalizing the sum of the individual prediction affinities to be at most one.

4. CONCLUSION

In this paper, we have presented a probabilistic method for predicting the binding of many TFs simultaneously. Our probabilistic method assumes that TFs compete for the binding on DNA, which mimics the situation in the cell. Our method provides a useful tool when predicting the regulating set of TFs. With the method realistic prediction results are achieved even though the set of TFs regulating a given gene is not known. To a large extent, this method overcomes the standard problems in TFBS prediction, such as a considerable amount of false positives and overlapping TFBSs.

Our method is also applicable in the situations when some knowledge of binding conditions is available. One can use in prediction only those TFs that are known to regulate the gene or those that are known to be present in the cell, for example results of protein or expression microarrays can be added to the model. At the moment one can add this information only by varying the set of TFs included in the prediction but in the future adding the TFs' concentration information to the model would give more accurate predictions. However, if some preferences of sequence positions are known (for example, if some position in sequence is known to be inaccessible), these can be added to the prior of the model as is done in [6]. With similar data integration one can integrate the knowledge of existing TF-DNA interactions or nucleosomes from ChIP-chip or ChIP-seq measurements as this kind of information improves the de novo motif discovery [11]. In the future, the existing protein-protein interactions will also be integrated into the model.

Our modeling results showed also some problems such as some of the binding sites were not found at all. We used TRANSFAC data to construct the priors for the PSFMs. Although our method implements Bayesian computation, prior information for some TFs can be too weak or even wrong. Besides, the quality of PSFMs varies significantly which can lead to weak binding affinities for some of the TFs. However, the binding affinity can be increased a lot by incorporating protein-protein interactions and by removing other interfering factors. To get more precise prediction results, more accurate knowledge of binding specificity in different conditions is needed. Such knowl-

Table 1. AUCs, when only binding for promoter matters.

False positive rate	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
Individual	0.0127	0.0253	0.0421	0.0651	0.0919	0.1214	0.1524	0.1865	0.2228
Individual normalized	0.0104	0.0237	0.0412	0.0644	0.0904	0.1194	0.1503	0.1845	0.2213
Competitive model	0.0185	0.0370	0.0606	0.0884	0.1179	0.1513	0.1849	0.2201	0.2578

edge can be achieved for example with protein binding microarray technique that covers all binding sites of a given length [12].

5. ACKNOWLEDGMENTS

Financial support from Tampere Graduate School in Information Science and Engineering (TISE) is gratefully acknowledged. Work was also supported by the Academy of Finland, (application number 129657, Finnish Programme for Centres of Excellence in Research 2006- 2011) and the Finnish Foundation for Technology Promotion.

6. REFERENCES

[1] S. Hannenhalli, "Eukaryotic transcription factor binding sites-modeling and integrative search methods," *Bioinformatics*, vol. 24, no. 11, pp. 1325–1331, Jun 2008.

[2] A. Wagner, "Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes," *Bioinformatics*, vol. 15, pp. 776–784, 1999.

[3] Z. Zhu, J. Shendure, and G. M. Church, "Discovering functional transcription factor combinations in the human cell cycle," *Genome Res*, vol. 15, pp. 848–855, 2005.

[4] A. Sinha, E. van Nimwegen, and E. D. Siggia, "A probabilistic method to detect regulatory modules," *Bioinformatics*, vol. 19, pp. i292–i301, 2003.

[5] N. Rajewsky, M. Vergassola, U. Gaul, and E. D. Siggia, "Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo," *BMC Bioinformatics*, vol. 3, pp. 30, Oct 2002.

[6] H. Lähdesmäki, A. G. Rust, and I. Shmulevich, "Probabilistic inference of transcription factor binding from multiple data sources.," *PLoS ONE*, vol. 3, Mar 2008.

[7] E. Blanco, D. Farré, M. M. Albà, X. Messeguer, and R. Guigó, "Abs: a database of annotated regulatory binding sites from orthologous promoters," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D63–67, Jan 2006.

[8] O. L. Griffith, S. B. Montgomery, B. Bernier, B. Chu, K. Kasaian, S. Aerts, S. Mahony, M. C. Sleumer, M. Bilenky, M. Haeussler, M. Griffith, S. M. Gallo, B. Giardine, B. Hooghe, P. V. Loo, E. Blanco, A. Ticoll, S. Lithwick, E. Portales-Casamar, I. J. Donaldson, G. Robertson, C. Wadelius, P. D. Bleser, D. Vlieghe, M. S. Halfon, W. Wasserman, R. Hardison, C. M. Bergman, and S. J. Jones, "Oreganno: an open-access community-driven resource for regulatory annotation.," *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D107–113, Jan 2008.

[9] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, P. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender, "Transfac: transcriptional regulation, from patterns to profiles," *Nucleic Acids Res*, vol. 31, pp. 374–378, 2003.

[10] M. M. Mason, Y. He, H. Chen, M. J. Quon, and M. Reitman, "Regulation of leptin promoter function by sp1, c/ebp, and a novel factor.," *Endocrinology*, vol. 139, no. 3, pp. 1013–1022, Mar 1998.

[11] L. Narlikar, R. Gordan, and A. J. Hartemink, "A nucleosome-guided map of transcription factor binding sites in yeast," *PLoS Comput Biol*, pp. 2199–2208, 2007.

[12] M. E. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. 3rd Estep, and M. Bulyk, "Compact, universal dna microarrays to comprehensively determine transcription factor binding site specificities.," *Nat Biotechnol*, vol. 24, no. 11, pp. 1429–1435, Nov 2006, bulyk.