Published online 28 September 2009

Nucleic Acids Research, 2009, Vol. 37, No. 22 ***e146***
doi:10.1093/nar/gkp789

A protein–protein interaction guided method for competitive transcription factor binding improves target predictions

Kirsti Laurila[1], Olli Yli-Harja[1] and Harri Lähdesmäki[1,2,*]

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

Received June 29, 2009; Revised August 30, 2009; Accepted September 7, 2009

ABSTRACT

An important milestone in revealing cells' functions is to build a comprehensive understanding of transcriptional regulation processes. These processes are largely regulated by transcription factors (TFs) binding to DNA sites. Several TF binding site (TFBS) prediction methods have been developed, but they usually model binding of a single TF at a time albeit few methods for predicting binding of multiple TFs also exist. In this article, we propose a probabilistic model that predicts binding of several TFs simultaneously. Our method explicitly models the competitive binding between TFs and uses the prior knowledge of existing protein–protein interactions (PPIs), which mimics the situation in the nucleus. Modeling DNA binding for multiple TFs improves the accuracy of binding site prediction remarkably when compared with other programs and the cases where individual binding prediction results of separate TFs have been combined. The traditional TFBS prediction methods usually predict overwhelming number of false positives. This lack of specificity is overcome remarkably with our competitive binding prediction method. In addition, previously unpredictable binding sites can be detected with the help of PPIs. Source codes are available at http://www.cs.tut.fi/~harrila/.

BACKGROUND

A significant proportion of cells' functions is determined by transcription of genes. Thus, it is important to understand the transcriptional regulation which is to a large extent controlled by transcription factors (TFs) binding to DNA. DNA sites that are bound by a TF can be identified by experimental methods, such as electromobility shift assay (EMSA). Moreover, recent high-throughput methods including chromatin immunoprecipitation-chip (ChIP-chip) or -sequencing (ChIP-seq) have increased our knowledge of the TF binding sites (TFBSs) remarkably. However, these experimental techniques are laborious and limited by the specificity of antibodies and additionally, they allow to study only one protein at a time in certain conditions. Hence, computational TFBS prediction methods have an important role in revealing genome-wide transcriptional regulation.

Most of the existing TFBS prediction methods consider the binding of a single TF at a time. These methods result in lot of false positive predictions as individual sequence motif models are sensitive but not very specific. Even though searching of all possible binding sites of one TF is important, it gives only a limited view of the whole transcription regulation processes of a cell. Rather than using only a single TF to regulate the expression of a gene, several TFs participate in the process in a combinatorial manner, in certain conditions and at the same time. Further, other DNA binding TFs are also present in the nucleus even though they may not regulate the gene of interest directly. If these TFs have accessible binding sites on the promoter of the studied gene, they can bind to DNA and block the binding of the other TFs. For example, in regulation of collagen type I (1) and in differentiation processes of hematopoietic stem cells (2), specific TFs can block the binding of other TFs that are participating in the regulation. Therefore, the transcription regulation process by TFs can be thought of as a competition between TFs. Those TFs that have the highest affinities to bind the sequence will, on average, win the competition of the binding site, but even those TFs that have lower affinities to this site have their

*To whom correspondence should be addressed. Email: kirsti.laurila@tut.fi
Correspondence may also be address to Harri Lähdesmäki Tel: +358 3 3115 11; Fax: +358 33 115 4989 Email: harri.lahdesmaki@tut.fi

© The Author(s) 2009. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

chance as determined by the steady state of the physical binding competition. Competition of binding sites is also affected by explicit interactions between regulatory TFs. For these reasons, studying the binding of all different TFs simultaneously is biologically more realistic than combining the predictions made for individual TFs.

A few schemas for predicting TFBS of multiple TFs at the same time already exist. These methods basically use two different approaches (3). The methods in the first category search for closely located binding sites as it is known that TFs interact with each other in the regulation process, and thus the TFBSs should be near to each other to allow interactions. These proximal TFBSs can then be applied to further searching and grouping to find regulating factors as has been done in (4–6). The other methods search for so-called *cis*-regulatory modules. These modules are clusters of binding sites for TFs that are known to affect expression together and to possibly interact with each other. Methods for searching *cis*-regulatory modules are presented, for example, in (7) where hidden Markov models and expectation maximization are used and in (8) which applies Gibbs sampler to the model.

In this article, we present a new method for predicting binding of several TFs simultaneously. Our method makes Bayesian inference for integrated probabilistic sequence specificity models and TFBSs and uses the prior knowledge of existing protein–protein interactions (PPIs) in prediction. Modeling results in a carefully constructed set of binding sites in the mouse genome show remarkable improvement compared with the cases where the individual prediction results of separate TFs have been combined. Especially the number of false binding sites is decreased significantly and previously unpredictable binding sites can be identified. A comparison with a widely used multiple TFBS prediction method, MSCAN (6), also shows the better performance of our model.

MATERIALS AND METHODS

MultiTF-PPI: a probabilistic model for competitive TF binding with PPIs

We formulate a PPI guided probabilistic model for competitive TF binding prediction, MultiTF-PPI. The goal of our method is to develop a biologically realistic model that mimics the situation in the cell. Thus, we take into account the existence of several TFs in the regulating process and their cooperation in the form of explicit and implicit interactions. As the knowledge of existing PPIs is not always available, we also provide a version of our multiple TF predictor without PPIs, MultiTF. In our modeling schema, we explicitly model simultaneous binding of several TFs to the same DNA sequence, which corresponds the situation where a large number of TFs compete for the binding to the same sites on a promoter. The proposed MultiTF-PPI method uses a similar idea as our previously developed probabilistic TF binding prediction method (9) which was developed for analyzing binding of a single TF together with additional sequence-level information. Here, we apply Bayesian inference to model binding of several TFs simultaneously as we know that practically all TFs can bind to any DNA sequence stretch, where the strength of binding is determined by the sequence affinity of each TF. Our probabilistic approach differs from the traditional use of deterministic binding sites model, such as position-specific frequency matrices (PSFMs), by modeling PSFMs as random variables and thereby providing a way to learn binding site characteristics from the data as well. In addition, as a result of using probabilistic framework, our method can detect both weak and strong TFBSs. This further justifies our modeling schema as also weak binding sites are known to be important for regulation processes (10). Furthermore, MultiTF-PPI provides all prediction results in terms of probabilities, which also allows us to answer quantitative questions of the TF binding. Most importantly, MultiTF-PPI also includes explicit interactions between TFs. The consequences of these interactions on TF binding are propagated into TF binding probabilities.

MultiTF-PPI is constructed in two steps: construction of the competitive binding model (MultiTF part), and incorporating interactions between regulatory TFs (PPI part). Our competitive method first models the binding of several TFs to a whole promoter sequence. Binding affinity of each TF is represented by a PSFM and nonbinding sites are represented by the standard Markovian background model. Using these definitions alone, it is straightforward to compute the probability of a sequence given the binding locations of a set of TFs. Instead of the standard frequentist computation, we implement a Bayesian alternative that integrates out the model parameters and computes the posterior probability of the given TFBSs. Computing Bayesian posterior probabilities for all TFBS locations and summing the probabilities of relevant locations gives a principled way of computing binding probabilities for any TF–TFBS pair or set of those pairs. 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 together with arbitrary complex PPIs.

Explicit PPIs are incorporated into MultiTF via an informative prior of TFBSs. The prior is constructed such that sufficiently closely located binding sites for TFs that have reported interactions are more probable *a priori*. We consider two different PPI priors: a piecewise flat prior where interacting TFs are assigned a higher probability as long as the TFBSs are within a length threshold, and a triangular prior where the prior strength is proportional to the distance of interacting TFs. The model version without interactions and preliminary results are also presented in our workshop report (11). See following sections for full details of the methods.

MultiTF: probabilistic model

Our probabilistic model for competitive TF binding prediction is built on the standard probabilistic building blocks. Let $X = (x_1,\ldots,x_L)$ denote a gene promoter, where $x_i \in \{\mathrm{A,C,G,T}\}$ and L is the length of the sequence. Sequence specificities of TFs are modeled using the PSFM. Let θ_k denote the ℓ_k-length PSFM for TF k, where $\theta_k(x_i,j)$ defines the probability of having nucleotide x_i at the j-th position of the PSFM. Note that the location within a PSFM $\theta_k(\cdot,j)$ is indexed as $j = 1,\ldots,\ell_k$. PSFMs for m TFs are collectively denoted by $\Theta = (\theta_1,\ldots,\theta_m)$. Nonbinding sites are modeled using the d-th order Markovian background model ϕ, where $\phi(x_i) = \phi(x_i|\mathrm{x}_{i\text{-}1},\ldots,x_{i\text{-}d})$ specifies the probability of observing nucleotide x_i given the previous d nucleotides. To model the binding of multiple TFs simultaneous, let $A = (a_1,\ldots,a_c)$ denote the start positions of c nonoverlapping binding sites on X, and vector $\pi = (\pi_1,\ldots,\pi_c)$ specifies which of the m TFs bind at each of the locations (i.e. $\pi_i \in \{1,\ldots,m\}$). Given these model assumptions, it is straightforward to compute the probability of sequence X as

$$P(X|A,\pi,\Theta,\phi) = P(X|\phi)\prod_{j=1}^{c} W_{a_j}^{(\pi_j)}, \tag{1}$$

where $P(X|\phi) = P(X|A = \emptyset,\phi) = \prod_{i=1}^{L}\phi(x_i)$ and

$$W_{a_j}^{(\pi_j)} = \begin{cases} \prod_{i=0}^{\ell_{\pi_j}-1} \frac{\theta_{\pi_j}(x_{a_j+i},i+1)}{\phi(x_{a_j+i})}, & \text{if } 1 \le a_j \le L-\ell_j+1, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Because binding specificities are known to contain a considerable amount of uncertainty, we let PSFMs as well as Markovian background model to be random variables and make Bayesian inference. We use Dirichlet priors for the PSWMs and the background model. Hyperparameters (pseudocounts) of the PSWMs are specified by the normalized TRANSFAC matrixes which is known to implement a strong regularization (12).

To make predictions of binding sites, a key quantity is the posterior probability that a set of TFs, defined by π, bind at specific locations defined by A, given a sequence X

$$P(A,\pi|X) \propto P(X|A,\pi)P(A,\pi). \tag{3}$$

The marginal likelihood is obtained by integrating over the parameters

$$P(X|A,\pi) = \int_{\Theta,\phi} P(X|A,\pi,\Theta,\phi)P(\Theta,\phi|A,\pi)\mathrm{d}\Theta\mathrm{d}\phi, \tag{4}$$

where the likelihood term is defined in Equations (1) and (2), and $P(\Theta,\phi|A,\pi) = P(\phi)\prod_{i=1}^{|A|}P(\theta_i)$ is the product of independent Dirichlet priors. Conjugate prior allows a closed-form solution to the marginal likelihood. Let n_{ijk} denote the number of times a nucleotide i is seen at the j-th position of the k-th PSFM (pseudocounts of the PSFMs α_{ijk} are defined analogously), and let $n_{ij}{}'$ denote the number of times a d-length nucleotide sequence j is followed by nucleotide i in the background model (pseudocounts of the background model $\alpha_{ij}{}'$ are defined analogously). Define $n_{jk} = \sum_i n_{ijk}$, $\alpha_{jk} = \sum_i \alpha_{ijk}$, $n'_j = \sum_i n'_{ij}$, and $\alpha'_j = \sum_i \alpha'_{ij}$. Equation (4) can be written as

$$P(X|A,\pi) = \prod_{j=1}^{4^d} \frac{\Gamma(\alpha'_j)}{\Gamma(\alpha'_j + n'_j)} \prod_{i\in\{\mathrm{A,C,G,T}\}} \frac{\Gamma(\alpha'_{ij} + n'_{ij})}{\Gamma(\alpha'_{ij})} \times \prod_{k=1}^{m}\prod_{j=1}^{\ell_k} \frac{\Gamma(\alpha_{jk})}{\Gamma(\alpha_{jk} + n_{jk})} \prod_{i\in\{\mathrm{A,C,G,T}\}} \frac{\Gamma(\alpha_{ijk} + n_{ijk})}{\Gamma(\alpha_{ijk})}, \tag{5}$$

For the basic method without PPIs, the prior $P(A,\pi)$ is defined as $P(A,\pi) = \sum_{c=0} P(A,\pi|c)P(c) = P(A,\pi|c = |A|)P(c = |A|)$, where c is again the number of TFBSs, $P(A,\pi|\mathrm{c})$ is uniform for a given c, and $P(c)$ is an exponentially decreasing function [see (9) for details].

MultiTF-PPI

Interactions between regulatory proteins are incorporated into the model via an informative prior $P(A,\pi)$, which probabilistically biases the search of TFBSs to those locations and configurations (A,π) where interacting TFs are within a close distance. Let $g(.)$ be a potential function that depends on the smallest distance between two TFBSs $y_{rs} = \min\{|\mathrm{a}_s + \ell_s - a_r|,|\mathrm{a}_s - a_r - \ell_r|\}$, $r \neq s$. We consider two potential functions, constant

$$g_c(y) = \begin{cases} w_1, & \text{if } y \le t_1, \\ 1 & \text{otherwise} \end{cases}, \tag{6}$$

and truncated triangular

$$g_t(y) = \begin{cases} w_2, & \text{if } y \le t_2, \\ \frac{w_2-1}{t_2-t_3}(y-t_3)+1, & \text{if } t_2 \le y \le t_3, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

The informative prior is defined as follows

$$P(A,\pi) \propto \prod_{r=1}^{c-1} g(y_{r(r+1)}), \tag{8}$$

where the empty product equals one and g is either g_c or g_t. In the analysis, we used parameters $t_1 = 80$, $w_1 = 3$, $t_2 = 60$, $t_3 = 12$ and $w_2 = 4$.

MCMC estimation

Direct computation of the posterior for all (A,π) is impossible. We develop a Metropolis–Hastings (MH) algorithm to sample from $P(A,\pi|X)$. The sampler is similar with the one proposed in (9) except that it is now extended to multiple TFs. Our MH implementation uses the following proposal distribution, $G(A',\pi'|A,\pi)$, to propose new TFBSs (A',π') given the current state (A,π):

- with probability p, for a randomly chosen TF, propose a new nonoccupied and nonoverlapping TFBSs uniformly randomly, if a free TFBS exists,
- with probability 1-p, delete a randomly chosen TFBS from (A,π), if a TFBS exists.

Proposed moves are accepted with the probability that satisfies the detailed balance condition

$$R = \min\left\{1, \frac{P(A',\pi'|X)}{P(A,\pi|X)} \frac{G(A,\pi|A',\pi')}{G(A',\pi'|A,\pi)}\right\}. \qquad \mathbf{9}$$

The sampler operates in a finite space and is easily seen to be ergodic by showing irreducibility and aperiodicity. Hence, the chain is guaranteed to converge to the correct posterior in the limit of infinite samples. After a burn-in period B, we collect a sample $((A^{(B+1)},\pi^{(B+1)}), (A^{(B+2)},\pi^{(B+2)}),\ldots,(A^{(B+N)},\pi^{(B+N)}))$. For finite samples, we monitor the convergence by running two parallel chains and measure the L_1-distance between binding probabilities of each TF over the whole promoter. The sample size N is increased as long as all the L_1 distances are below a threshold (1×10^{-4}).

Statistical inference

We are interested in computing two types of posterior probabilities, the probability that TF k binds the location j, and the probability that TF k binds somewhere in the promoter. Define $\mathcal{A}_{jk} = \{(A,\pi) : \exists i : a_i = j \wedge \pi_i = k\}$. The first quantity can then be easily computed

$$P(\text{'TF } k \text{ binds at location } j \text{ of sequence } X\text{'}) = \sum_{(A,\pi)\in\mathcal{A}_{jk}} P(A,\pi|X). \qquad \mathbf{10}$$

The second binding probability can be computed similarly by defining $\mathcal{A}_k = \{(A,\pi) : \exists i : \pi_i = k\}$ and summing

$$P(\text{'TF } k \text{ binds the promoter sequence } X\text{'}) = \sum_{(A,\pi)\in\mathcal{A}_k} P(A,\pi|X). \qquad \mathbf{11}$$

A useful alternative to Equation 11 is to take the maximum of the binding probabilities over a promoter, i.e.

$$P(\text{'TF } k \text{ binds the promoter sequence } X\text{'}) \approx \max_j \sum_{(A,\pi)\in\mathcal{A}_{jk}} P(A,\pi|X). \qquad \mathbf{12}$$

Note that, despite the max operation, Equation (12) includes all the aspects of competitive binding and PPIs. The above probabilities can be directly estimated from the MCMC chain using sample averages which converge to true posterior probabilities almost surely.

Data

We validated our results with the test set used in (9) that was originally generated from ABS (13) and OregAnno (14) databases (test data set is publicly available at http://www.probtf.org/). We filtered the primary test set by removing sequences (and binding sites) for which nonredundant PSFMs for the TFs were not known. After this we had 29 promoter sequences. Promoter sequences were cut 50 nt before the first known TFBS and 50 nt after the last one. The lengths of sequences varied between 110 nt and 1322 nt.

Nonredundant set of PSFMs used in this study was collected from TRANSFAC [(15); Release 10.3]. After this, 27 different TFs had binding sites in the test set and this set of TRANSFAC matrices was used in all predictions. Markov model parameters were estimated from a set of 250 upstream noncoding sequences that are believed not to contain TFBSs.

Used PPIs were collected from the Biomolecular Interaction Network Database [BIND; (16)], the Biological General Repository for Interaction Datasets [BioGRID; (17)] and the Human Protein Preference Database [HPRD; (18)]. Only interactions for animals/mammalians were used, depending on the databases possibilities. As the same protein can have many different aliases, these were searched from GenBank (19) to avoid excluding the existing interactions.

RESULTS

Competitive binding modeling reduces false positives

We tested the MultiTF by computing the TFBS predictions for our test set. These predictions were compared with the results where individual predictions for each TF (9) were combined. Individual predictions for TFs using the method from (9) without any additional data is known to perform slightly better than the widely applied standard scanning methods for binding site prediction (20–22) and hence represents a valuable point of comparison. Overall, the results of MultiTF show substantial improvement when compared with the combined individual predictions. The individual predictions often result in physically impossible situations where several TFBSs are overlapping, which is illustrated in Figures 1 and 2. In the first figure the three known TFBSs for mouse leptin promoter are shown. TF Sp1 is known to bind to region −100/−95 of the promoter (relative to transcription starting site), cEBP to region −58/−28 and TBP to the region −33/−28 (23). However, other TFs could also bind to the promoter, especially MYB, Sp1 and TEAD have high affinities to the cEBP binding site as demonstrated in Figure 2.

As an example case we predicted TFBSs for the mouse leptin promoter region −145/−1 with MultiTF, MultiTF-PPI and combined individual predictions and the results of these predictions are shown in Figures 3 and 4. In the first figure, only the predictions for three TFs that are

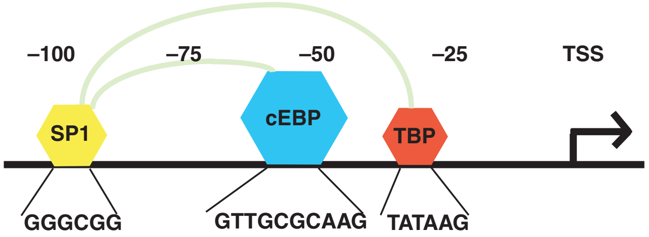

Figure 1. Known binding sites for mouse leptin (U36238) promoter. Interactions known to exist (from databases) between TFs are marked with shaded curves. TSS denotes the transcription starting site.

known to bind the promoter, Sp1, cEBP and TBP are shown and in the other figure the predictions for all 27 TFs used in the analysis are shown. For Sp1, the binding site at −100/−95 was correctly predicted with all of the methods. However, with individual combined prediction, analysis resulted in four false binding site predictions for Sp1, whereas with MultiTF only one false prediction was made. The individual combined predictions also resulted to the situation where the cEBP binding site have four different candidate TFs to bind, cEBP (the true one), MYB, SP1 and TEAD (Figure 4a). On the other binding sites the situation was similar. Thus, our results suggest that the commonly used combined individual prediction approach to TFBS prediction generates a large number of false positives and complicated overlapping results, with increasing severity when the number of TFs is increased. With MultiTF, this problem is clearly diminished as the number of false positives is reduced from 10 to 2 (in addition, in both predictions there always exist some very weak binding sites). Thus, MultiTF overcomes the main problem of traditional methods, which is the number of false positives. Furthermore, with MultiTF the binding of cEBP was predicted correctly instead of choosing some of the other three false TFs. The known binding site for TF TBP on leptin promoter could not be predicted with either of the methods, which stems most probably from TBP's incompatible binding specificity model. The results for the rest of the data set were similar as combined individual predictions led to several overlapping and contradicting predictions and MultiTF predicted far less binding sites. A more comprehensive comparison of results is shown in the next sections.

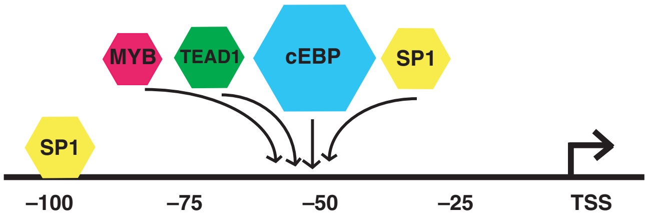

Figure 2. TFBS binding site predictions when individual predictions are combined. The SP1 binding site is predicted correctly but for the cEBP binding site predictions suggest also binding for MYB, SP1 and TEAD. Only predictions on these two binding sites are shown.

Interactions between regulatory TFs strengthen TFBS predictions

We compared MultiTF-PPI with the other methods after adding an informative prior for TFBSs constructed from

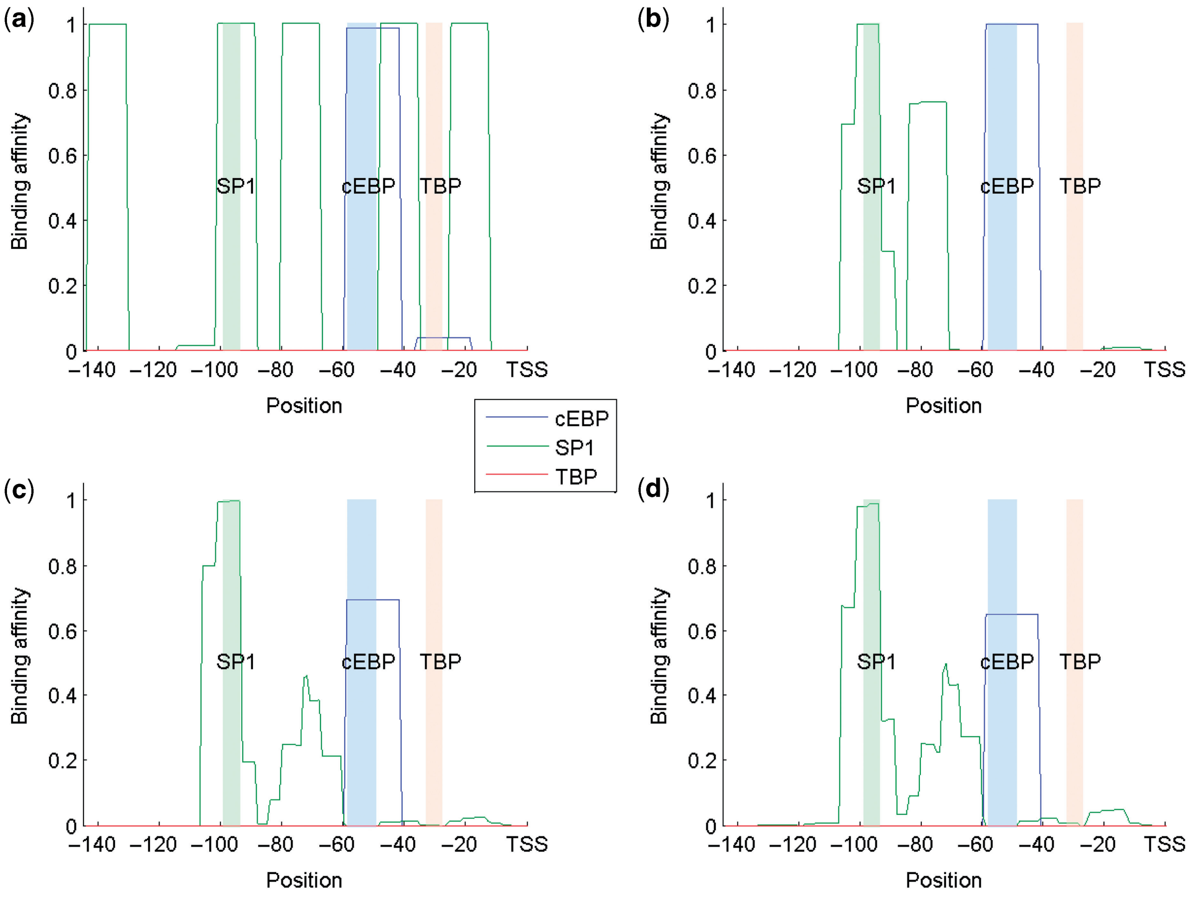

Figure 3. Predictions for mouse leptin promoter. Predictions are presented only for those TFs that are known to bind to the promoter. Known binding sites are shaded. (**a**) Individual predictions, (**b**) MultiTF, (**c**) MultiTF-PPI, piecewise flat prior, and (**d**) MultiTF-PPI, triangular prior.

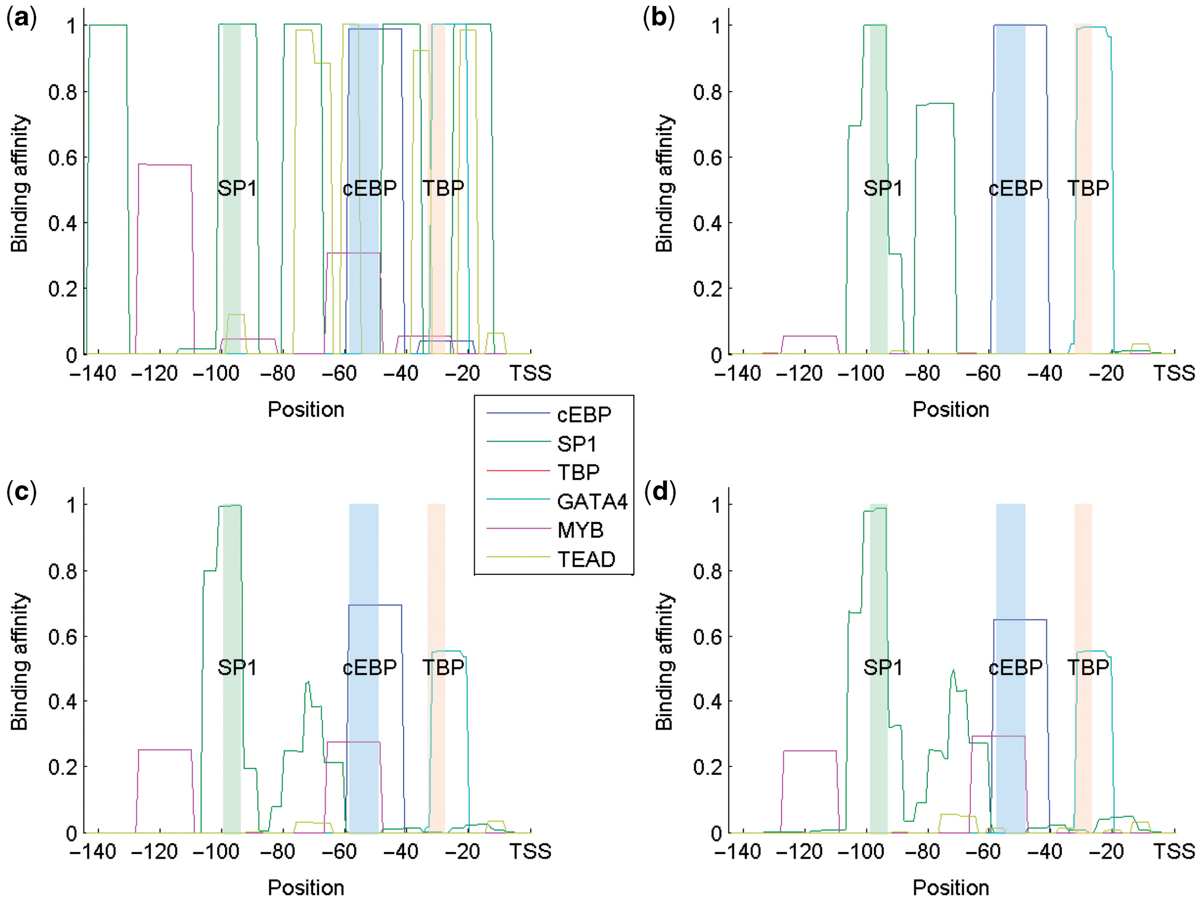

Figure 4. Predictions for mouse leptin promoter. Predictions are presented for those TFs that bind with affinity more than 0.1. Known binding sites are shaded. (**a**) Individual predictions, (**b**) MultiTF, (**c**) MultiTF-PPI, piecewise flat prior and (**d**) MultiTF-PPI, triangular prior.

PPIs. We used two different priors, the piecewise prior and the triangular prior. For the MultiTF-PPI with piecewise prior, we increased the prior with a factor of three if the binding sites were closer than 80 nt and for the triangular prior, we started from the weight four with the distance 12 and weight strength dropped linearly to the one when distance between binding sites were 60 or more. These weights and distances gave the best results when testing different values for the parameters (for weights we tested parameters between two and five and for distances values between 10 and 100). The distance between piecewise prior is ~7.6 DNA turns and 5.7 for the triangular prior, so in both models there exist several proximal major and minor grooves which the interacting TFs can use for binding. In general, with piecewise flat prior the best results were obtained with weight three with varying distance from 40 to 90 nt ,whereas with triangular prior the starting weights four and five with maximum distance for weighting from 40 to 60 nt had the best performance.

The prediction results with MultiTF-PPI gave generally similar results compared with MultiTF (e.g. Figure 3c and d and Figures 4c and d). However, in several cases, interactions improved predictions with strengthening the binding affinity of a real binding site or by weakening the affinity of a false one. Producing predictions with MultiTF may result for finding only weak binding signals as can be seen in Figure 5, which illustrates the TFBSs predictions for the acetylcholinesthերase (*ACHE*) promoter. Promoter has six annotated binding sites for three different TFs, of which the first two adjacent Sp1 binding sites compete with the EGR1 of binding (see the shaded truncated binding sites in the Figure 5) (24). First, note the enormous amount of false positive predictions for the combination of individual predictions in Figure 5a. Again it is indeterminable which of the overlapping predictions are the right ones as apart from Sp1, also GATA4, TEAD1 and HNF4 are predicted to bind on to the promoter with a high affinity. With MultiTF (Figure 5b) majority of false predictions are diminished totally or partly. However, even though the number of false positives is dramatically reduced, only the first two binding sites of Sp1 are predicted clearly, whereas the third shows only a weak binding site. When the prediction is run with MultiTF-PPI, the situation changes. The method incorporates for the binding model the effect of known homotypic interaction of Sp1 and the heterotypic interaction between Sp1 and EGR1. Prediction results show (Figure 5c) that with piecewise flat prior both of the Sp1 binding sites are predicted correctly. Nonetheless, use of the interactions raises the amount of

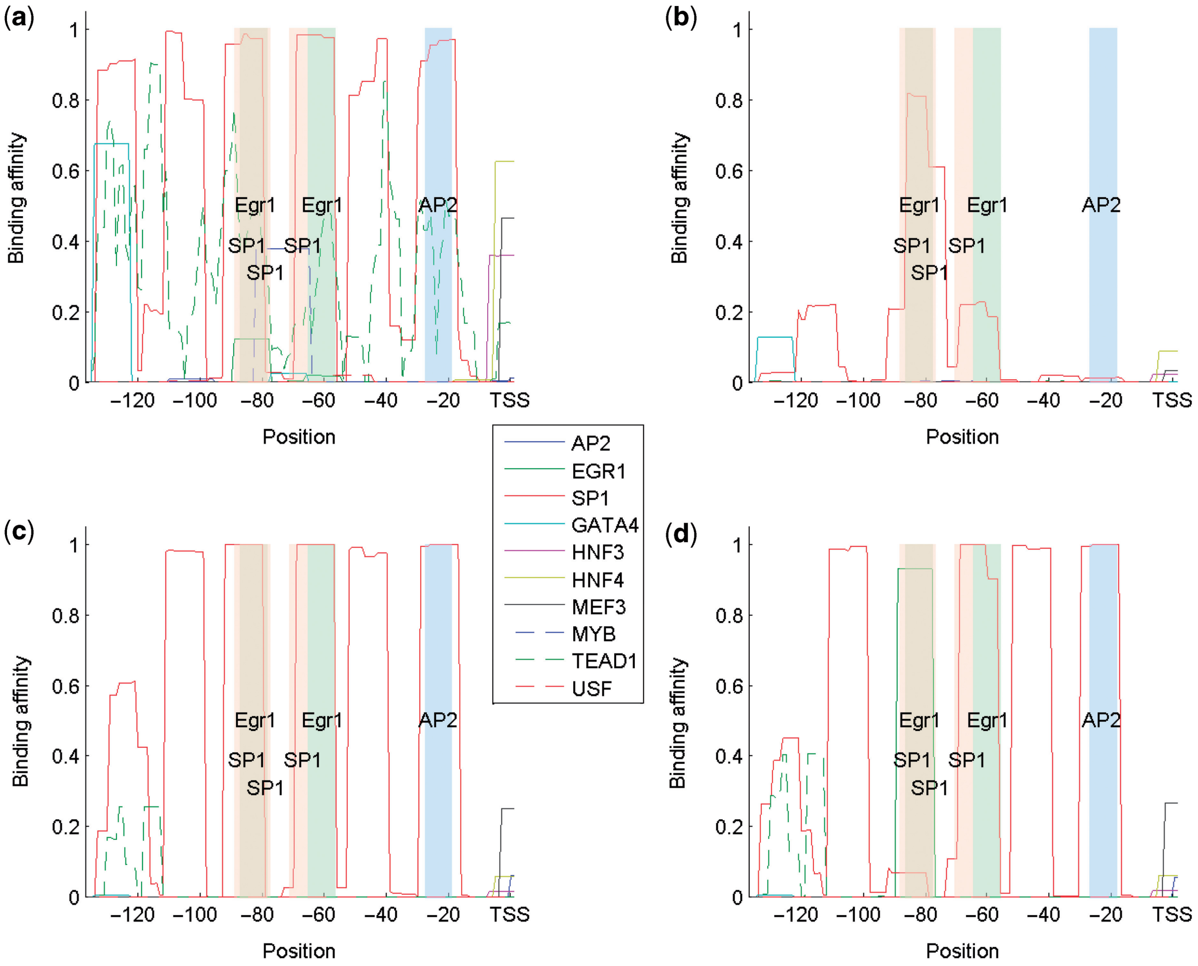

Figure 5. Predictions for mouse acetylcholinesteterase promoter. Predictions are presented for those TFs that bind with affinity more than 0.1. Known binding sites are shaded. (**a**) Individual predictions, (**b**) MultiTF, (**c**) MultiTF-PPI, piecewise flat prior and (**d**) MultiTF-PPI, triangular prior.

false predictions of that of the MultiTF. However, the amount of false positives is still far from the number of that obtained with combined individual predictions. Furthermore, if one is interested only in the regulating set of TFs, the false positives of MultiTF-PPI are not interfering the result. When we perform the prediction using triangular prior, the situation changes a little. The Sp1 binding is predicted similarly as with the piecewise flat prior except the first binding sites. Instead of predicting two first Sp1 sites, we have predicted EGR1 binding site which could not be predicted with the other method versions. As Sp1 and EGR1 really compete for on this binding site in the cell, the competitive modeling schema as such can mimic the biological regulation process.

Incorporating PPIs into the competitive binding model does not only reduce the false positive rate (FPR) but also can lead to finding of new binding sites that cannot be predicted with other models. An example of this can be seen in Figure 6 where the predictions and real binding sites for the activating transcription factor 2 (*ATF2*, preciously known also as *CREB2*) promoter are shown. The promoter has five different binding sites for Sp1 which is known to form homocomplexes. With combined individual predictions, only one Sp1 binding site can be predicted and further, there exists several strong binding sites for eight other TFs. With MultiTF the false predictions are reduced to only three additional TFs, but the single predicted Sp1 binding site is very weak. When the MultiTF-PPI is used, three of the five Sp1 binding sites are predicted correctly. Neither individual predictions nor MultiTF could predict these binding sites at all. The false predictions of the PPI model are extra occurrences of Sp1 and no other TFs have strong false affinity to the promoter unlike with individual combined or MultiTF prediction methods have.

Competitive methods improve TFBS predictions

Above, we have showed concrete examples of how our models ameliorate the TFBS predictions. We also evaluated the overall performance of our methods to identify TF target genes with receiver operating characteristic (ROC) curves and the area under the curve (AUC). The ROC curves in Figures 7 and 8 [obtained with Equations (12) and (11), respectively] show large improvement in the performance of all the competitive methods relative to the combined individual predictions.

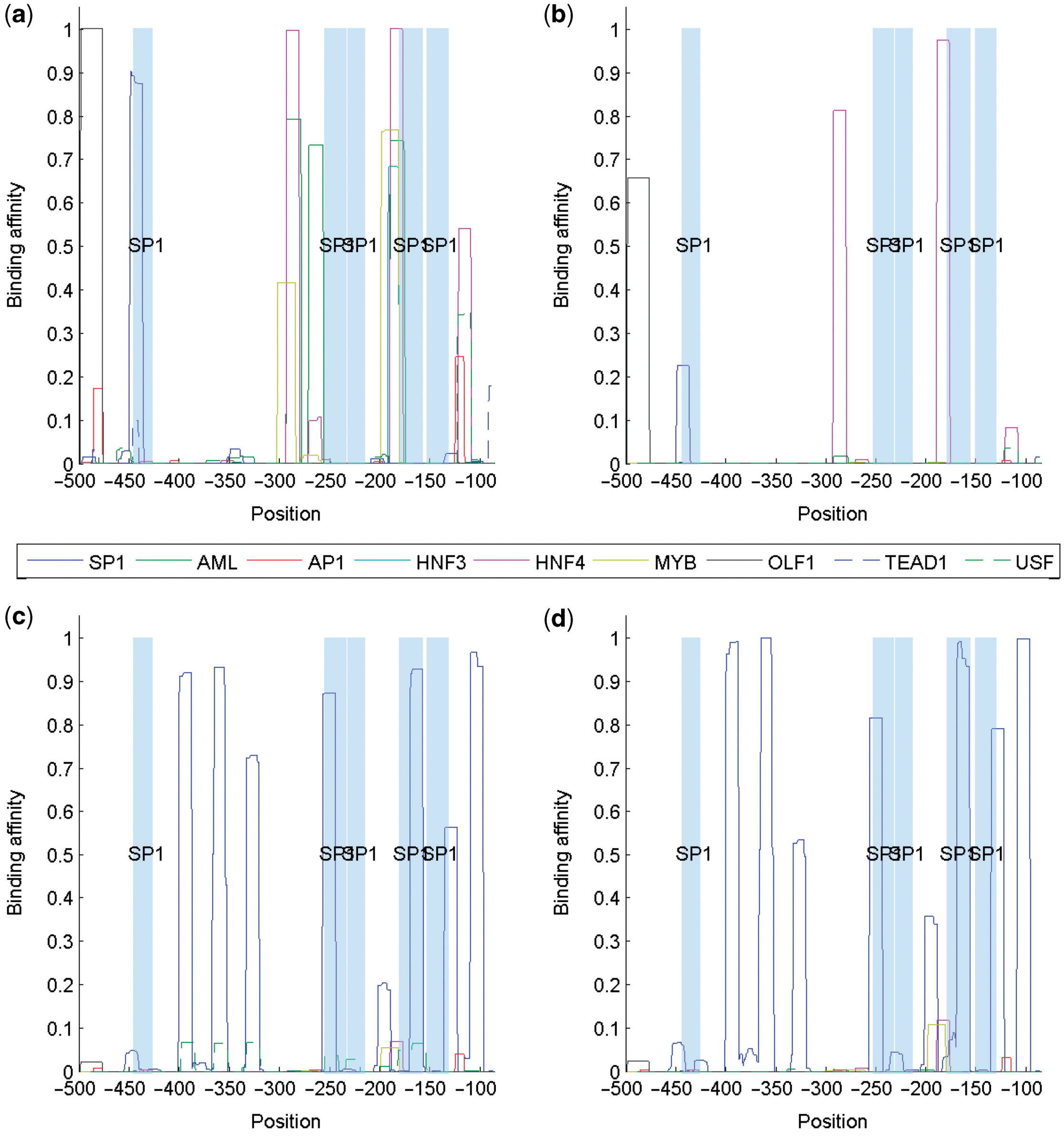

Figure 6. Predictions for mouse activating transcription factor 2 promoter. Predictions are presented for those TFs that bind with affinity more than 0.1. Known binding sites are shaded. (**a**) Individual predictions, (**b**) MultiTF, (**c**) MultiTF-PPI, piecewise flat prior and (**d**) MultiTF-PPI, triangular prior.

In Figure 7, after the competitive binding modeling with or without PPIs, binding estimates are obtained by computing the maximum binding probability over the promoter sequence, whereas in Figure 8 binding estimates are computed using the total binding probability. Based on these ROC curves, the maximum binding probability approach (computed from MultiTF-PPI output) seems to give comparable or slightly better discrimination performance than the total probability approach. This better discrimination is possibly due to low specificity of the PSFM models, which, therefore, can result in several weak false positive binding site predictions. Although such weak false predictions have a low probability, summing over the whole promoter can decrease the relative importance of the real, high-probability binding sites and thus increase the background noise. The biggest improvement is found with small and most preferable FPRs. Competitive models are far more discriminative from the random case than the individual predictions. The same result can be found as the AUC scores [for the Equation (11)] in Table 1, where we also report the AUC scores for the case where the sum individual predictions are normalized to be one at most in a single binding site. The normalization, however, weakens the power of individual predictions and thus we have compared our methods with the nonnormalized case. For FPR 0.1, the AUC of the MultiTF-PPI with both of the priors is twice as large as the AUC of combined individual predictions. Further, the AUC of the sole MultiTF improves the individual predictions with >45% higher AUC score. In total, the MultiTF-PPI with triangular prior shows to be slightly better than the version with piecewise flat prior. The MultiTF-PPI improves the simple MultiTF when FPR is <0.3 and thus is favorable especially for appropriate, small FPR values.

We also tested the effect of randomly generated interaction data on predictions. Random PPIs were generated by preserving the number of PPIs in the original data (see details from Supplementary Data). Supplementary Figures S1 and S2 show the expected behavior that MultiTF-PPI with random PPIs performs worse than with the correct PPIs. Surprisingly, MultiTF-PPI with random PPIs seems to perform slightly better than MultiTF without interactions. Because the PPI adjacency matrix is relatively dense in our case, on average one-fourth of the randomly generated PPIs were also correct ones. The slight improvement in the results compared with the MultiTF predictions can be understood by the fact that the proposed method incorporates PPIs probabilistically by biasing (but not explicitly forcing) binding sites of interacting proteins close to each other. Overall, this suggests that the method is not sensitive to noisy interaction information.

We also compared our results with those of MSCAN, method that searches functional TFBS clusters. We run the program with default parameters and changed only the minimum number of hits to one to allow also the occurrence of single binding sites in the predictions. With MSCAN, we achieved TPR 0.0250 with FPR 0.0042. When we compared our results with the other methods both the prior forms of MultiTF-PPI outweigh MSCAN clearly, with piecewise flat prior TPR being 0.0909 and with triangular prior true positive rate (TPR) being 0.0649. With MultiTF, the nearest FPR point is at 0.0057 where TPR is 0.0390. However, MSCAN performs better than normalized combined individual predictions: with FPR of 0.0042, TPR is only 0.0130.

The running time of the algorithm depends on several parameters, including promoter sequence length and the number of TFs. The running time also depends on a user-tunable parameter for the MCMC estimation convergence threshold, for which we have used a rather stringent threshold of 1×10^{-4}. In our simulations, typical running times are of the order of 30 min per promoter. In applications, reasonably well-converged results can be obtained with much less stringent convergence thresholds, say 1×10^{-2}, which decreases the running time considerably.

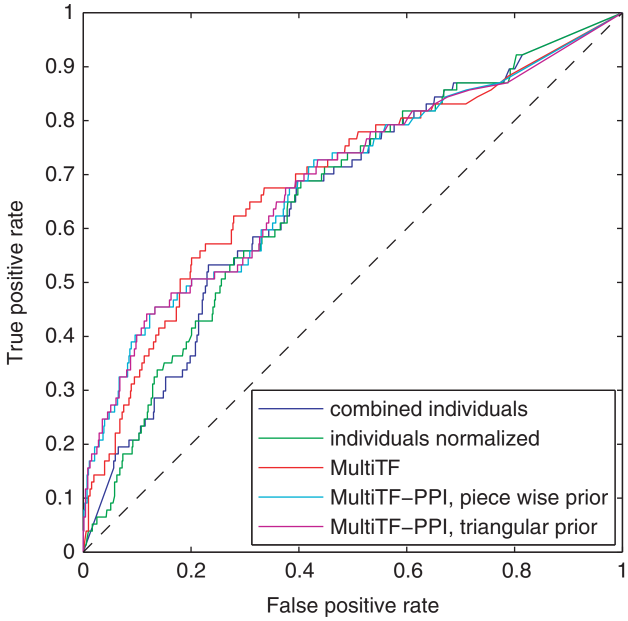

Figure 7. ROC curves for different methods [computed with Equation (12)].

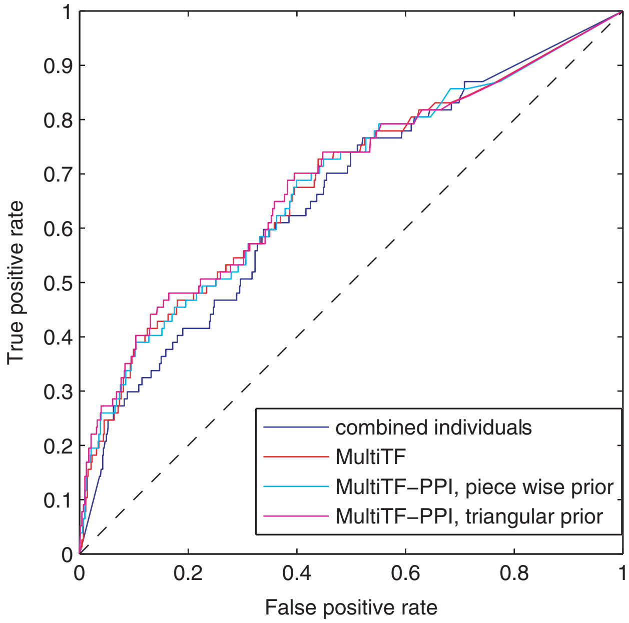

Figure 8. ROC curves for different methods [computed with Equation (11)].

Table 1. AUC scores for different methods

FPR	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
MultiTF	0.0185	0.0370	0.0606	0.0884	0.1179	0.1513	0.1849	0.2201	0.2578
MultiTF-PPI, piecewise flat prior	0.0252	0.0465	0.0705	0.0957	0.1215	0.1505	0.1826	0.2168	0.2554
MultiTF-PPI, triangular prior	0.0252	0.0467	0.0709	0.0961	0.1220	0.1514	0.1841	0.2192	0.2556

DISCUSSION

In this article, we have presented a probabilistic method, MultiTF-PPI that predicts the binding of many TFs simultaneously. Our method is based on biologically realistic assumptions and uses probabilistic framework to model the competition for the binding of several TFs on the promoters. We have added the known PPIs to the model as a prior knowledge to improve predictions. This model can increase the binding affinity of a TF that interacts with already bound TFs or prevents displacing a bound TF that interacts with other TFs. We also present a method without PPIs. Both of our proposed methods are useful tools when predicting the TFBSs on DNA sequences. Methods can be used both in the situations where no prior knowledge of regulating factors exists and in the cases when the regulating set (or part of it) for TFs of a particular gene is already known. In both the cases, realistic prediction results are obtained as the well-known problems in traditional TFBS prediction, such as high number of false positives and overlapping TFBSs, are overcome.

If the set of regulating TFs is unknown, our method allows to use a large set of possible regulators in the prediction. Additionally, interfering TFs can also be added to the model to better mimic the transcription process in nucleus. Even though a set of several regulating proteins is used, the number of predicted TFBSs remains reasonable, which is usually not the case when predicting the binding sites of TFs separately. In the situations where some information about of regulating TFs is available, our methods can also be applied with those TFs that are known to regulate the gene. This allows the use of knowledge of existing protein data obtained with mass spectrometry-derived methods or with the other quantitative measurements' along the same lines as has been done in (10). Further, other additional data sources can be added to the model as a prior knowledge as illustrated in (9). This kind of valuable prior knowledge can be obtained, for example, from nucleosome locations (25) and conservation data (26), which are shown to improve motif discovering and TFBS predictions. These additional information sources are easily incorporated and combined into our model.

The use of our competitive TFBS methods allows also to find TFBSs that cannot be found by individual predictors. This overcomes some of the problems that are encountered while using PSFMs as a binding model. PSFMs have been built using a varying number of verified binding sites, which results to varying quality of the matrices. Thus, many TFs can be predicted to have only a weak affinity on DNA. These binding affinities can be, nevertheless, strengthened by using PPIs.

Besides the quality of PSFMs, also the characteristics of the PPI databases can increase the noise level of the predictions. Many PPIs can be found from proteome-wide databases, but most of the interactions are reported to take place in regulation process of certain genes in certain conditions. PPIs may also require a specific order of interacting proteins or the interactions may exist only with a fixed distance, albeit bending of DNA allows some flexibility on this. These different preferences of distances between different PPIs would be useful to incorporate into the model. Another source of uncertainty is caused by complexes that are formed by several TFs. In these complexes all of the TFs do not bind to DNA at all. We tried to integrate this information to the model, too, but when we allowed interactions also via one transmitter protein, the modeling results deteriorated. Nonetheless, this result was quite expected, as the noise in the PPI databases cumulates and can affect the results even more than with the direct PPIs. As part of future work, we will study if information about known or predicted protein complexes can improve the prediction accuracy. The third problem of using PPI data is the lack of knowledge whether some factors can interact at all. However, it is currently impossible to gather comprehensive information about noninteracting proteins as all the possible conditions and gene regulation processes cannot be tested experimentally.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Funding for open access charge: Tampere Graduate School in Information Science and Engineering (TISE); Academy of Finland (application number 213462, Finnish Programme for Centres of Excellence in Research 2006-2011); Finnish Foundation for Technology Promotion.

Conflict of interest statement. None declared.

REFERENCES

1. Ponticos,M., Partridge,T., Black,C.M., Abraham,D.J. and Bou-Gharios,G. (2004) Regulation of collagen type I in vascular smooth muscle cells by competition between Nkx2.5 and δ EF1/ZEB1. *Mol. Cell Biol.*, **24**, 6151–6161.
2. Koschmieder,S., Rosenbauer,R., Steidl,U., Owens,B.M. and Tenen,D.G. (2005) Role of transcription factors C/EBPalpha and PU.1 in normal hematopoiesis and leukemia. *Int. J. Hematol.*, **81**, 386–377.
3. Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites-modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.
4. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
5. Zhu,Z., Shendure,J. and Church,G.M. (2005) Discovering functional transcription factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.
6. Alkema,W.B.L., Johansson,O., Lagergren,J. and Wasserman,W.W. (2004) MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, **32**, W195–W198.
7. Sinha,A., vanNimwegen,E. and Siggia,E.D. (2003) A probabilistic method to detect regulatory modules. *Bioinformatics*, **19**, i292–i301.
8. Rajewsky,N., Vergassola,M., Gaul,U. and Siggia,E.D. (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. *BMC Bioinformatics*, **3**, 30.

9. Lähdesmäki,H., Rust,A.G. and Shmulevich,I. (2008) Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, **3**, e1820.
10. Segal,E., Raveh-Sadka,T., Schroeder,M., Unnerstall,U. and Gaul,U. (2008) Predicting expression patterns from regulatory sequence in *Drosophila* segmentation. *Nature*, **451**, 535–540.
11. Laurila,K. and Lähdesmäki,H. (2009) A probabilistic model for competitive binding of transcription factors. *In Proceedings of the Sixth TICSP Workshop on Computational Systems Biology (WCSB 2009)*. Aarhus, Denmark, pp. 107–110.
12. Steck,H. and Jaakkola,T.S. (2002) On the Dirichlet prior and Bayesian regulation. *In Advances in Neural Information Processing Systems*. Vol. 15. MIT Press Cambridge, MA, pp. 697–704.
13. Blanco,E., Farré,D., Albà,M.M., Messeguer,X. and Guigó,R. (2006) ABS: a database of annotated regulatory binding sites from orthologous promoters. *Nucleic Acids Res.*, **34**, D63–D67.
14. Griffith,O.L., Montgomery,S.B., Bernier,B., Chu,B., Kasaian,K., Aerts,S., Mahony,S., Sleumer,M.C., Bilenky,M., Haeussler,M. *et al.* (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res.*, **36**, D107–D113.
15. Matys,V., Fricke,E., Geffers,R., Gossling,E., Haubrock,M., Hehl,R., Hornischer,K., Karas,D., Kel,A.E., Kel-Margoulis,O.V. *et al.* (2003) TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
16. Alfarano,C., Andrade,C.E., Anthony,K., Bahroos,N., Bajec,M., Bantoft,K., Betel,D., Bobechko,B., Boutilier,K., Burgess,E. *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.*, **33**, D418–D424.
17. Stark,C., Breitkreutz,B.J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
18. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R., Raghavan,T.M. *et al.* (2006) Human protein reference database-2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
19. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2008) GenBank. *Nucleic Acids Res.*, **36**, D25–D30.
20. Staden,R. (1984) Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.*, **12**, 505–519.
21. Quandt,K., Frechl,K., Karas,H., Wingender,E. and Werner,T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878–4884.
22. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
23. Mason,M.M., He,Y., Chen,H., Quon,M.J. and Reitman,M. (1998) Regulation of leptin promoter function by Sp1, C/EBP, and a novel factor. *Endocrinology*, **139**, 1013–1022.
24. Mutero,A., Camp,S. and Taylor,P. (1995) Promoter elements of the mouse acetylcholinesterase gene transcriptional regulation during muscle differentiation. *J. Biol. Chem.*, **270**, 1866–1872.
25. Narlikar,L., Gordan,R. and Hartemink,A.J. (2007) A nucleosome-guided map of transcription factor binding sites in yeast. *PLoS Comput. Biol.*, **3**, e215.
26. Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **338**, 276–287.