Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 235795, 15 pages
doi:10.1155/2010/235795

Research Article

Novel Data Fusion Method and Exploration of Multiple Information Sources for Transcription Factor Target Gene Prediction

Xiaofeng Dai,[1,2] Olli Yli-Harja,[1] and Harri Lähdesmäki[1,3]

[1] *Department of Signal Processing, Tampere University of Technology, P.O. Box 553, 33101 Tampere, Finland*
[2] *Institute of Molecular Medicine, University of Helsinki, P.O. Box 20, 00014 Helsinki, Finland*
[3] *Department of Information and Computer Science, Aalto University School of Science and Technology, P.O. Box 15400, 00076 Aalto, Finland*

Correspondence should be addressed to Xiaofeng Dai, xiaofeng.dai@helsinki.fi and Harri Lähdesmäki, harri.lahdesmaki@tut.fi

Received 17 April 2010; Revised 29 June 2010; Accepted 10 August 2010

Academic Editor: Byung-Jun Yoon

Background. Revealing protein-DNA interactions is a key problem in understanding transcriptional regulation at mechanistic level. Computational methods have an important role in predicting transcription factor target gene genomewide. Multiple data fusion provides a natural way to improve transcription factor target gene predictions because sequence specificities alone are not sufficient to accurately predict transcription factor binding sites. *Methods*. Here we develop a new data fusion method to combine multiple genome-level data sources and study the extent to which DNA duplex stability and nucleosome positioning information, either alone or in combination with other data sources, can improve the prediction of transcription factor target gene. *Results*. Results on a carefully constructed test set of verified binding sites in mouse genome demonstrate that our new multiple data fusion method can reduce false positive rates, and that DNA duplex stability and nucleosome occupation data can improve the accuracy of transcription factor target gene predictions, especially when combined with other genome-level data sources. Cross-validation and other randomization tests confirm the predictive performance of our method. Our results also show that nonredundant data sources provide the most efficient data fusion.

1. Introduction

A central problem in molecular system biology is to understand the manner in which a cell operates its complex transcriptional machinery. At molecular level, transcriptional processes are largely controlled by transcription factors (TFs) that bind to gene promoters in a sequence-specific manner and, thereby, inhibit or promote the expression of their target genes. Collectively, these DNA-binding proteins and other molecules work together to implement the complex regulatory machinery that controls gene expression. Since large-scale understanding of transcriptional regulation is still severely limited even in lower organisms, it is of great importance to reveal these regulatory protein-DNA interactions genomewide.

Experimentally verified TF-binding sites (TFBSs) have been collected in databases [3–5] and recently developed experimental methods, such as ChIP-chip or ChIP-seq, are capable of measuring *in vivo* TFBSs in high-throughput manner. However, it is not possible to obtain sufficient coverage, that is, to screen all TFs under all conditions, using experimental methods alone. Therefore, the binding site prediction problem calls for computational methods. Computational predictions rely on sequence specificities that are typically taken from a database [4] or obtained as an output from a motif discovery method [6]. Recent progress on experimental side has made it also possible to measure TF-binding specificities in high-throughput manner [7]. The advent of these experimental techniques equips TF target gene prediction methods with much more accurate binding specificity models and, indeed, opens a whole new avenue for computational analysis of TF-DNA binding.

Sequence specificities alone, however, are not sufficiently informative to accurately predict TFBSs simply because

the probability of observing an exact copy of a presumably functional binding motif in a genome by chance is remarkably high. A natural way to improve TF target gene predictions is to incorporate additional information into statistical inference of TFBSs. A number of additional data sources can be useful for this purpose, including, among others, information on coregulated genes, evolutionary conservation, physical binding locations as measured by ChIP-chip or ChIP-seq, nucleosome occupancies, CpG islands, regulatory potential, DNase hypersensitive sites, and so on. Incorporating additional information sources to guide statistical inference has successfully been made use of in the context of motif discovery [8–11], but has not attracted enough attention in TF target gene prediction. We have recently developed a probabilistic TF target gene prediction method, ProbTF, which can incorporate practically any additional genome-level information source to predict TF target gene [12].

Statistical data fusion for TF target gene prediction becomes more challenging in the case of multiple information sources. Here we develop a new method for multiple data fusion and incorporate novel data sources into TF target gene prediction. Four genome-level additional information sources (i.e., information at the level of individual nucleotides), evolutionary conservation, nucleosome positioning data from a recently published computational method, regulatory potential, and DNA duplex stability, are employed here to improve TF target gene prediction, which is expected to be informative of binding sites as will be discussed shortly. Some of these and other individual data sources have already been shown to improve *de novo* motif discovery [8–11]. Here we demonstrate how multiple data sources can be combined to make joint statistical inference of TF target gene. Integration of data sources that have a probabilistic interpretation is relatively straightforward [12], and for other data sources we convert the raw data into probabilities, or prior distributions, by extending a previously proposed Bayesian transformation method [11]. In addition, for efficient use of DNA duplex stability data, we propose a simple heuristic that can assess the binding preference (single versus double-stranded DNA) for a TF from a set of known binding sites. Results on a carefully constructed set of verified binding sites in mouse genome [3, 5, 12] demonstrate that the new data fusion method that we propose here improves the performance of TF target gene prediction methods. We also demonstrate that a number of genome-level data sources, either alone or especially in combination, are highly informative of TF target gene. Consequently, our statistical data fusion method can gain valuable new insights into genomewide models of transcriptional regulatory networks.

2. Methods

Given the fundamental role of TFs in transcriptional regulation, we focus on predicting TF target gene. Because each individual data source is noisy and gives only a partial view of the underlying regulatory mechanisms, we focus on making statistical inference for TFBSs from multiple information sources. The essence of the data fusion problem that we encounter is illustrated in Figure 1, which shows four examples of verified binding sites from the test data set together with the associated additional genome-level data sources [12]. The first row in each subplot shows the annotated binding site(s) for a TF in a gene promoter. The next rows (named by their TRANSFAC IDs, grey) show the log-likelihood scores of the position specific frequency matrix (PSFM) models to the Markovian background model ϕ. The following five rows show the additional data sources: probability of conservation (con. [13], green), regulatory potential (reg. [14], blue), nucleosome positioning signals predicted by two different methods (npy. [1] and nuc. [2], magenta), and DNA duplex stability (DNA. [15, 16], red) score for each position of the sequences. The joint prior combined from all the explored additional data sources is shown in black in the last row. The median and mean of the scores for each data type applied to the sequences shown in Figure 1 are recorded in Table S1 in supplementary material available online at doi: 10.1155/2010/235795.

Figure 1 shows that the highest log-likelihood score is not always obtained at the annotated binding site. TFs are commonly associated with multiple PSFMs since one TF may allow certain variation in its binding motif. Thus, it can be difficult to combine predictions from multiple PSFMs given that these PSFMs may be extremely similar or different. This issue can be solved by, for example, ProbTF method, which implements an intuitive way of combining predictions by multiple PSFMs: ProbTF considers all possible numbers of nonoverlapping TFBSs in all possible locations and configurations and weights each configuration according to its probability. A more difficult problem is to decide that which of the peaks predicted by PSFMs correspond to real, functional binding sites. As illustrated in Figure 1, the PSFM-based profiles have relatively good sensitivity but poor specificity, which is common to many PSFMs. The lack of specificity can be greatly improved by genome-level data fusion, which forms the focus of this study.

Corresponding to what is known about transcriptional regulation, many of the verified binding sites typically have high degree of conservation [8] and high regulatory potential scores [14] and are typically free of stable nucleosomes (i.e., have low nucleosome occupancy scores) [17]. Moreover, DNA double helix destabilization energies at TF binding sites are different from those at random sites [11]. In particular, TFBSs tend to have high DNA duplex stability score if a TF prefers to bind both strands of the promoter sequence (Figures 1(a) and 1(b)) and low DNA stability score in the opposite case (Figures 1(c) and 1(d)) The above reasoning seems to provide a simple logic for filtering the real TFBSs.

However, correlation between TFBSs and any of the additional data sources cannot be expected to be perfect even from a biological point of view. For example, only about 50% of functional binding sites are assessed to be evolutionary conserved [18]. The additional information sources are also noisy, regardless of whether they are experimental measurements or computational predictions. The only possibility is to make statistical inference, which takes the inherent

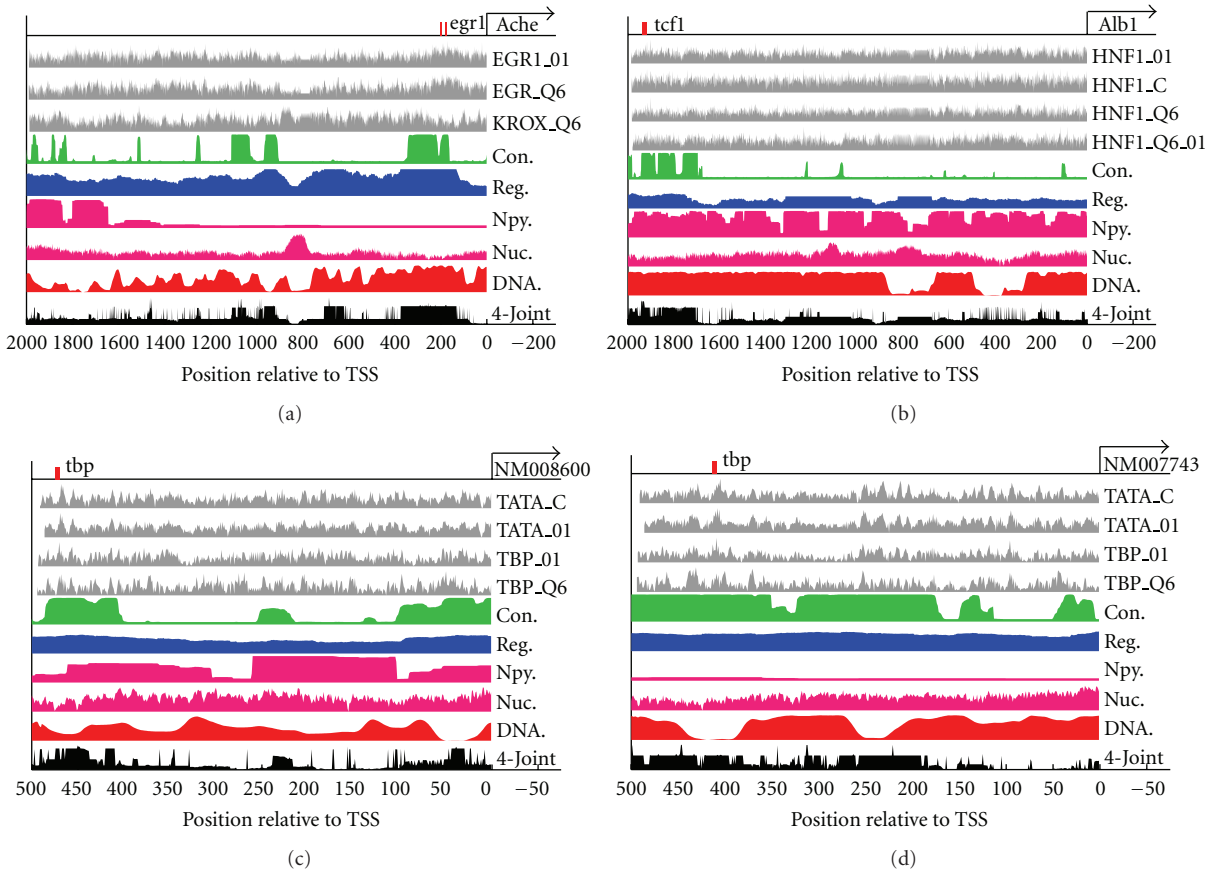

FIGURE 1: *Illustration of data fusion problem in TF target gene prediction.* The promoter sequence names are shown above the arrow, and the arrow corresponds to transcription start site (TSS). Horizontal axis corresponds to position relative to TSS. The red bar(s) together with a TF name on the first line of each figure represent the known binding site. For a given TF, data shown in grey (named with TRANSFAC IDs) represent models corresponding to different position-specific frequency matrices (PSFM) that are found from the TRANSFAC database. Evolutionary conservation (green), regulatory potential (blue), two nucleosome positioning signals [1, 2] (magenta), and DNA duplex stability data (red) are shown in the following five rows (abbreviated with con., reg., npy., nuc. and DNA., resp.). The joint prior from all the four additional data sources (black) is shown in the last row. TFs shown in panels (a) and (b) are assumed to bind to their corresponding sequences in a double-strand manner, while TFs in panels (c) and (d) bind in a single-strand manner. All plotted data are for mouse genome.

randomness into account, from multiple genome-level data sources. The rationale is that the accuracy of computational TF target gene predictions naturally improves when more (useful) information is incorporated into statistical analysis.

2.1. Probabilistic Framework for TF Target Gene Prediction. We first describe the TF target gene prediction algorithm employed in this study (full details can be found from [12]). Let $S = (s_1, \ldots, s_N)$ denote a single strand of a promoter sequence, where $s_i \in \{A, C, G, T\}$ and N is the length of the sequence (generalization to double-stranded DNA sequences is also possible but omitted here). Let Q denote the number of (unknown) binding sites and A the (hidden) start positions of nonoverlapping binding sites in sequence S; that is, if $Q = c$ then $A = \{a_1, \ldots, a_c\}$.

Nonbinding site (i.e., background) sequence locations are modeled by the dth order Markovian background model ϕ. Assuming that we have access to the d previous nucleotides before the start of the actual sequence S, the likelihood of a sequence S having no binding sites for any TF is $P(S \mid A = \varnothing, \phi) = P(S \mid \phi) = \prod_{i=1}^{N} \phi(s_i)$, where $\phi(s_i) = P(s_i \mid s_{i-1}, \ldots, s_{i-d})$. We set $d = 0$ since that value provides the best results in [12]. TFBSs are modeled with the standard PSFM model which is a product of independent multinomial distributions. Let $\theta(s_i, j) = P_\theta(s_i, j)$ denote the probability of observing nucleotide s_i at the jth ($j = 1, \ldots, l$) position of θ, where l is the length of the motif. Assume a TF is characterized by M PSFMs, $\Theta = (\theta^{(1)}, \ldots, \theta^{(M)})$. Define $\pi \in \{1, \ldots, M\}^c$ as the configuration of motif models from Θ in A; that is, π_i specifies the motif model $\theta^{(\pi_i)}$, which begins from location a_i and has a length l_{π_i}. Further, the probability of sequence S, given nonoverlapping motif positions and the motif and background models, is

$$P(S \mid A, \pi, \Theta, \phi) = P(S \mid \phi) \prod_{j=1}^{|A|} W_{a_j}^{\pi_j}, \quad (1)$$

where $|A| = Q = c$ and

$$W_{a_j}^{\pi_j} = \begin{cases} \prod_{k=0}^{l_{\pi_j}-1} \dfrac{\theta^{(\pi_j)}\left(s_{a_j+k}, k+1\right)}{\phi\left(s_{a_j+k}\right)}, & \text{if } 1 \le a_j \le N - l_{\pi_j} + 1 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The probability that a sequence S has c binding sites is obtained with Bayes' rule

$$P(Q = c \mid S, \Theta, \phi) = \frac{P(S \mid Q = c, \Theta, \phi) P(Q = c \mid \Theta, \phi)}{P(S \mid \Theta, \phi)}, \tag{3}$$

where the normalization factor is $P(S \mid \Theta, \phi) = \sum_{c=0}^{\lfloor N/l_{\min} \rfloor} P(S \mid Q = c, \Theta, \phi) P(Q = c \mid \Theta, \phi)$ and $\lfloor N/l_{\min} \rfloor$ is the maximum number of nonoverlapping motifs in an N-length sequence. As proposed in [12], the prior of the number of motif instances, $P(Q = c \mid \Theta, \phi)$, is assumed to be independent of Θ and ϕ and has an exponential form

$$P(Q = c) \sim \left[\frac{1}{2}, \frac{1}{C}, \frac{\kappa}{C}, \frac{\kappa^2}{C}, \ldots, \frac{\kappa^{\lfloor N/l_{\min} - 1 \rfloor}}{C} \right], \tag{4}$$

where $C = 2 \sum_{i=0}^{\lfloor N/l_{\min} \rfloor - 1} \kappa^i$. We use $\kappa = 0.5$. This formula defines the (user definable) prior expectation of the number of binding sites in a given DNA sequence. Importantly, it does not incorporate any of the informative data sources studied here. This prior, primarily only, increases or decreases of the estimated binding probabilities, and as such has little effect on, for example, the ROC curves. The probability $P(S \mid Q = c, \Theta, \phi)$ is obtained with the assumption that, for a fixed value of Q, the prior over binding site positions A and configurations π is uniform and inversely proportional to the number of different binding site positions and configurations. The probability $P(S \mid Q = c, \Theta, \phi)$ is obtained by summing over all possible positions and configurations, and can be computed efficiently using a recursive formula [12].

Finally, the probability that a TF which is characterized by Θ binds to a promoter S, $P(\Theta \rightarrow S \mid S, \Theta, \phi)$, is defined as the probability that at least one of the motif models in Θ has a binding site in S

$$P(\Theta \longrightarrow S \mid S, \Theta, \phi) = P(Q > 0 \mid S, \Theta, \phi). \tag{5}$$

Integration of additional data sources into the aforementioned probabilistic TF target gene prediction framework is carried out by assuming that the data sources are in the form of $D = (P_1, \ldots, P_N)$ where P_i is the probability that the ith base pair location is a binding site. D can be derived from a single data source or from multiple data sources (see subsections "DNA duplex stability data", "Nucleosome occupation data", and "Data integration method" of this section for details). Assuming that S and D are conditionally independent and the probability of D does not depend on the PSFM and background models, the probability of S and D given A, π, Θ, and ϕ is

$$P(S, D \mid A, \pi, \Theta, \phi) = P(S \mid A, \pi, \Theta, \phi) P(D \mid A, \pi). \tag{6}$$

Following (1), the probability $P(D \mid A, \pi)$ is modeled as

$$P(D \mid A, \pi) = \prod_{i=1}^{N} (1 - P_i) \prod_{j=1}^{|A|} \prod_{k=0}^{l_{\pi_j}-1} \frac{P_{a_j+k}}{1 - P_{a_j+k}}, \tag{7}$$

and, thus, the joint probability $P(S, D \mid A, \pi, \Theta, \phi)$ can be written compactly as

$$P(S, D \mid A, \pi, \Theta, \phi) = P(S \mid \phi) P(D \mid \phi) \prod_{j=1}^{|A|} W_{a_j}^{(\pi_j)} \times D_{a_j}^{(\pi_j)}, \tag{8}$$

where $P(D \mid \phi) = \prod_{i=1}^{N} (1 - P_i)$ and $D_{a_j}^{(\pi_j)} = \prod_{k=0}^{l_{\pi_j}-1} ((P_{a_j+k})/(1 - P_{a_j+k}))$. Consequently, the same efficient recursive algorithm can be used to compute $P(\Theta \rightarrow S \mid S, D, \Theta, \phi)$ (see [12] for more details).

Note that the choice of Markovian and PSFM models is arbitrary. Also note that since additional data are incorporated using probabilities of binding over the promoter sequence; we could also employ methods other than ProbTF.

2.2. Data Integration Method. Define the mth additional genome-level data source (for a single gene promoter having length N) as $D^{(m)} = (\mathcal{P}_1^{(m)}, \ldots, \mathcal{P}_N^{(m)})$, $1 \le m \le n$. Denote the probabilities for position i from n different data sources as $\mathcal{P}_i = (\mathcal{P}_i^{(1)}, \ldots, \mathcal{P}_i^{(n)})$, $1 \le i \le N$. Further, define a thresholded version of probabilities $\mathcal{P}_i^{(m)}$ as

$$\widetilde{\mathcal{P}}_i^{(m)} = \begin{cases} \mathcal{P}_i^{(m)}, & \text{if } \mathcal{P}_i^{(m)} \ge T^{(m)}, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $T^{(m)}$ is a threshold for the mth data source and is defined as a percentile q of the distribution of the mth data source. Then the thresholded scores for position i can be written as $\widetilde{\mathcal{P}}_i = (\widetilde{\mathcal{P}}_i^{(1)}, \ldots, \widetilde{\mathcal{P}}_i^{(n)})$, $1 \le i \le N$. Let $v_i = |\{\widetilde{\mathcal{P}}_i^{(m)} \mid \widetilde{\mathcal{P}}_i^{(m)} > 0, 1 \le m \le n\}|$ be the number of data sources that exceed their thresholds at location i, then the integrated probability for position i, $\widetilde{P}_i$, is calculated as

$$\widetilde{P}_i = \begin{cases} \max\left(\widetilde{\mathcal{P}}_i\right) \times L_{v_i}, & \text{if } v_i \ge 1, \\ \min(\mathcal{P}_i) \times L_0, & \text{otherwise.} \end{cases} \tag{10}$$

The data integration method is parameterized by $L_0, L_1, \ldots, L_n$ and q. Note that $v_i \in \{0, 1, \ldots, n\}$ and $L_{v_i+1} \ge L_{v_i}$. It is also worth noting that the resulting probabilities do not include hard thresholding for any of the genomic locations although thresholding is involved in integration, and the use of thresholding during the construction is motivated by its simple yet powerful parametrization.

The data integration method is illustrated in Figure 2 for the case of two additional data sources with parameters $L_0 = 0.5$, $L_1 = 0.7$, $L_2 = 1$, and $q = 0.9$. For illustration purposes, both data sources are assumed to have uniform distribution and hence $T^{(1)} = T^{(2)} = q$.

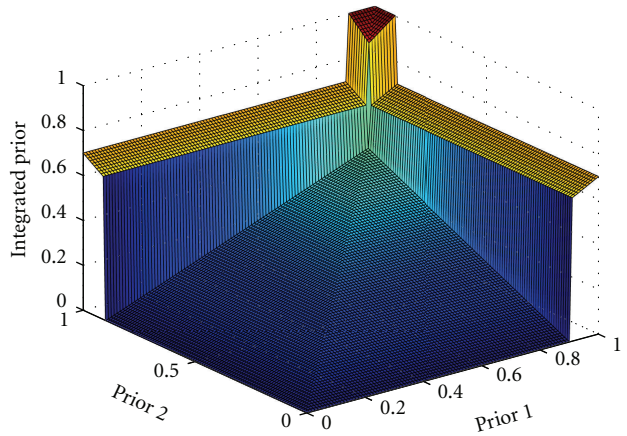

Figure 2: *An illustration of the prior integration method.* An illustration of the prior integration method for the case of two additional data sources. x and y axes correspond to the two data sources and z-axis corresponds to the integrated prior.

In the above genome-level data integration method there are $n + 1$ (n is the number of additional data sources) weighting parameters $L_0, L_1, \ldots, L_n$, and one threshold q for emphasizing the most informative binding locations. There are also two scaling parameters, a multiplicative factor $a^{(m)}$, and a bias term $b^{(m)}$, for each additional data source, and one scaling parameter, c, for combining other data sources with the TF target gene prediction analysis. These parameters are used to scale the original probability values into a proper range. In particular, the scaling parameters are used in the following way (for the mth data source):

$$\mathcal{P}_i^{(m)} = a^{(m)} \times P\left(\mathcal{X} \in \mathbf{B} \mid R^{(m)}(\mathcal{X})\right) + b^{(m)},$$
$$P_i = 2 \times c \times \widetilde{P}_i + 0.5 - c, \tag{11}$$

where $P(\mathcal{X} \in \mathbf{B} \mid R^{(m)}(\mathcal{X}))$ is the probability that a DNA site $\mathcal{X}$ is a TFBS ($\mathcal{X} \in \mathbf{B}$) given the value of the mth raw data $R^{(m)}(\mathcal{X})$. For conservation and regulatory potential the original data are already in a probabilistic format, and for nucleosome and DNA stability data the conversion of the raw data into probabilities was described in the previous sections. Probability P_i is the final integrated prior probability for position i after scaling, which is directly used in further TFBS prediction as explained, for example, in (6) and (7).

All the parameters needed in this study were chosen by a grid search method via optimizing receiver operating characteristic (ROC) curves, and the importance of each data source could be reflected by the multiplicative factor "a"; that is, the higher the multiplicative factor the less noisy or more important this type of data is. "1-specificity" (x axis) and "sensitivity" (y axis) are used to draw the ROC curves according to

$$\text{specificity} = \frac{\text{TN}}{(\text{FP} + \text{TN})},$$
$$\text{sensitivity} = \frac{\text{TP}}{(\text{TP} + \text{FN})}, \tag{12}$$

where TN, FP, TP, FN each stands for "true negative", "false positive", "true positive", and "false negative", respectively. In particular, TN, FP, TP, FN are obtained by comparing the computed binding probabilities (of a sequence to have a binding site for a TF) with known binding site information from the test data set, that is, "0" (no binding site) and "1" (at least one binding site). We used area under the curve (AUC) and AUC30 (the AUC for the area between false positive rates $[0, 0.3]$) to optimize the parameters. In case of four additional data sources, we are dealing with a high-dimensional grid search problem. Since the grid size grows exponentially with the dimension, we resort to a heuristic where each parameter is optimized separately using a 1-dimensional grid search while keeping the other parameters fixed. Moreover, parameter optimization is done sequentially so that we first optimize parameters $a^{(m)}$ and $b^{(m)}$ for individual data sources. Scaling parameters $L_0, L_1, \ldots, L_n$ are optimized similarly except that L_n is always assigned to 1. For example, parameters L_1 and L_0 are optimized using two data sources, which are then kept fixed and assigned to L_2 and L_1, respectively, when optimizing new parameter L_0 using three data sources, so forth. In our study, we optimized the parameters of up to four data sources, which are $L_0 = 0.72$, $L_1 = 0.72$, $L_2 = 0.73$, $L_3 = 0.8$, and $L_4 = 1$, respectively, and q equals 0.93. It is worth noticing that the adjacent L_w's ($0 \le w \le n$) tend to be similar for small values of w, and especially we have $L_1 = L_0$ when n equals 4. This accords well with the main feature of our new data fusion method, which is to search for *bona fide* locations (indicated by several data sources) and reduce false positives by not paying too much attention to the locations indicated by fewer data sources. All the rest optimized scaling parameters are listed in Table 1.

The scaling parameters, that is, "a", "b" and "c", are relatively robust, whose slight variations would not dramatically affect the results. We varied "a" of the DNA duplex stability data (for both double and single strand binding data), which is supposed to have more effect on the results (recall that "a" weights different information sources and reflects their importance), and listed its AUC scores for single data source as well as its combination with other additional information sources in supplementary Table S2. It is clear that with small changes of "a", the results do not vary significantly. However, for the weighting parameters, that is, "L_0" to "L_n", and the threshold, q, their small changes may have greater effect on the results, since they determine how different data sources are combined. This can be seen from the closer values among "L_0" to "L_3", which are 0.72, 0.72, and 0.73, respectively. These parameters depend heavily on the quality and type of data, and should be optimized before data integration.

2.3. DNA Duplex Stability Data. The DNA stability measures the amount of energy needed to separate the two strands of DNA. In this study the DNA destabilization energies were obtained from an online tool WebSIDD [15, 16], where the parameters were set to "DNA Type: circular", "Energetic Type: near neighbor", and "Energy Cutoff: level 4". Note that circular DNA is assumed to calculate the duplex stabilities of linear DNA. This is because WebSIDD handles linear DNA

similarly with circular DNA but adding 50 G/C to the end, which is not needed here given the extended DNA used. We obtained the energy score for each sequence with 1 kb extension from both ends. For every binding site $\mathcal{X}$ we computed the energy of destabilization $G(\mathcal{X})$ as the average of the destabilization values $G(\mathcal{X}, i)$ for all positions i within this site.

2.3.1. Assessing Binding Preference for Each TF. Relatively little is reported about specific types of protein-DNA interactions in the literature and the protein domain annotations are not available for all TFs, thus, we decided to assess the binding preference for each factor simply by looking at the DNA stability scores at the known binding sites in the test data set. With the assumption that the binding preference of a TF is the same to all its binding sites, we estimated the binding preference of each TF with the following heuristic. Let $\mathcal{A}$ denote the set of all known start binding positions of a TF among all the tested sequences in our test set. For all the known binding sites in $\mathcal{A}$, we compute counts dC and sC which are the number of times $\sum_{i=a_j}^{a_j+\ell_j-1} G(\mathcal{X}, i)/\ell_j \geq T$ and $\sum_{i=a_j}^{a_j+\ell_j-1} G(\mathcal{X}, i)/\ell_j \leq 1 - T$, respectively, where ℓ_j is the width of the verified binding site j in the test set and T is the threshold specified by quantile q. Then, the TF is assigned to bind in a double-strand manner if $dC > sC$, in a single-strand manner if $dC < sC$, and in cases $dC = sC$, random preference is assigned. In order to make the above heuristic more robust, we repeated it for three thresholds specified by different quantiles $\mathbf{q} = \{0.6, 0.7, 0.8\}$ with both raw $G(\mathcal{X}, i)$ and smoothed $\overline{G}(\mathcal{X}, i) = \sum_{j=i}^{i+9} G(\mathcal{X}, j)/10$ DNA duplex stability scores. The final binding preference of each TF is made by taking a vote among these six binding preferences, and again in case of a tie random binding preference is assigned.

2.3.2. Construction of DNA Duplex Stability Prior. We built three data sets to construct the DNA duplex stability priors: one positive single-strand binding data set, one positive double-strand binding data set, and one background data set. The positive data sets are constructed from 226 known binding sites in our test data set by splitting the known binding sites into single- and double-strand binding sets according to the binding preference of each TF. The background data set is generated as follows. For each verified binding site in our test set, we randomly select 20 genomic locations (from the same promoter sequence) with the average binding site of length 12, which results in a background set that is 20 times larger than the test set.

The raw DNA duplex stability scores are converted into probabilities using a similar method as in [11] with an extension to account for both single- and double-strand binding preferences. For each data set, we built a histogram of the energies, then normalized and smoothed the values to get a probability distribution. The cumulative distribution functions (CDFs) of the three data sets are shown in Figure 3(a), which indicate that DNA duplex stability data does provide us discriminative information about TFBSs. All known binding sites, on which the performance is eventually evaluated, are used to draw Figure 3(a), which leads to circular reasoning. However, our cross-validation and randomization simulations show that this biasing effect is negligible. For every energy value e and binding site $\mathcal{X}$, the conditional density of the single- and double-strand binding data are $P(G(\mathcal{X}) = e \mid \mathcal{X} \in s\mathbf{B})$ and $P(G(\mathcal{X}) = e \mid \mathcal{X} \in d\mathbf{B})$, respectively, where $s\mathbf{B}$ and $d\mathbf{B}$ denote single- and double-strand TFBSs, respectively. Similarly, for the random genomic locations we have $P(G(\mathcal{X}) = e)$. We also estimated the frequency of the randomly chosen DNA sites that have a significant overlap with any of the known single-strand and double-strand binding sites, $P(\mathcal{X} \in s\mathbf{B})$ and $P(\mathcal{X} \in d\mathbf{B})$, respectively. Bayes' rule is used to compute the probability that a DNA site $\mathcal{X}$ is a single-strand TFBS given its energy (similar calculation is also applied to the double-strand case)

$$P(\mathcal{X} \in s\mathbf{B} \mid G(\mathcal{X})) = \frac{P(G(\mathcal{X}) \mid \mathcal{X} \in s\mathbf{B}) \times P(\mathcal{X} \in s\mathbf{B})}{P(G(\mathcal{X}))}. \tag{13}$$

2.4. Nucleosome Occupation Data

2.4.1. Construction of Nucleosome Occupation Prior. We built the nucleosome occupation prior in a similar way as what we did with the DNA stability data, but with only two data sets: positive and background (see also [11]). The positive data set consists of the averaged N-scores (the raw nucleosome occupancy scores obtained using the method in [2]) of the known binding positions. The background data set is composed of the averaged N-scores of randomly selected genomic locations in the same way as above. For every occupation score o, the conditional probabilities for binding and nonbinding sites are denoted as $P(N(\mathcal{X}) = o \mid \mathcal{X} \in \mathbf{B})$ and $P(N(\mathcal{X}) = o)$, respectively. The CDFs of the two nucleosome data sets are shown in Figure 3(b), which indicate that the nucleosome positioning information from [2] is informative of TFBSs. The probability that a DNA site $\mathcal{X}$ is a TFBS given its nucleosome occupation score is obtained by (13) (with $s\mathbf{B}$ replaced by $\mathbf{B}$). Note that $P(\mathcal{X} \in \mathbf{B}) = P(\mathcal{X} \in s\mathbf{B}) + P(\mathcal{X} \in d\mathbf{B})$.

2.5. Data.

We validate our computational methods using the same mouse data set as in [12], which consists of 47 promoter sequences (as shown in Table 2), each with a varying number of annotated binding sites from ABS [3] and ORegAnno [5] databases (the annotated binding sites are also listed in Table 2). Sequence lengths are 2 Kbps or vary around 500 bps. PSFM models are taken from TRANSFAC [4] (professional version 10.2). The additional data sources used here are conservation, regulatory potential, DNA duplex stability, and nucleosome positioning. The first two data sources are the same as what have been used in [12], where conservation is assessed with the PastCons scores [13] and regulatory potential is constructed from a set of known regulatory and nonregulatory sequences using a discriminatory computational analysis (prediction algorithm is named "ESPERR") [14]. DNA duplex stability, and nucleosome positioning are the two new data sources explored in more detail in this study. We use our computational methods to

Table 1: *AUC scores and scaling parameters for all data sources and their combinations.* Data source combinations from 0 to 4 information sources are colored grey, green, blue, yellow, and magenta, respectively. "a" and "b" are the multiplicative factor and bias term, respectively, for scaling each additional data source, and "c" is the scaling parameter used for combining multiple information sources into the TF target gene prediction framework. All the parameters shown here are selected with respect to the largest AUC scores.

Data combination	a	b	c	Auc	AUC (CV)
npy.	0.01	0.49		0.6986	
no prior				0.7449	
nuc. + DNA.			0.12	0.7501	
nuc.	0.04	0.45		0.7555	0.7464
DNA.	0.06 (s), 0.01 (d)	0.49		0.7580	0.7484
reg.	0.1	0.45		0.7611	0.7465
reg. + nuc. + DNA.			0.06	0.7741	
reg. + DNA.			0.05	0.7771	
reg. + nuc.			0.06	0.7946	
con.	0.08	0.46		0.8038	0.7492
con. + reg. + DNA.			0.06	0.8143	
con. + reg. + nuc. + DNA.			0.07	0.8154	
con. + DNA.			0.07	0.8174	
con. + reg.			0.06	0.8220	
con. + nuc. + DNA.			0.09	0.8253	
con. + reg. + nuc.			0.06	0.8284	
con. + nuc.			0.07	0.8334	

predict that whether the promoter of a gene has TFBS(s) or not.

3. Results and Discussion

In this section, the results of exploring two novel additional data sources, evaluating the new data fusion method and comparison among different data source combinations in TF target gene prediction are sequentially reported and discussed. The idea of our computational methods is to probabilistically bias the search of binding sites to those genomic locations that are more likely to contain binding site(s) in light of the additional data. The qualities of the TF target gene prediction results are evaluated by the ROC curves and the histograms of the estimated binding probabilities, which are drawn from the probabilities over all the TFs and the sequences being analyzed. The test data set used throughout this paper consists of 47 promoter sequences, each contains a varying number of annotated binding sites from ABS [3] and ORegAnno [5] databases.

3.1. Novel Informative Data Sources

3.1.1. DNA Duplex Stability Prior. Most sequence-specific DNA binding proteins contact with the major groove of double stranded DNA in the B conformation [19], and some TFs are shown to bind DNA in a double-strand manner according to their crystal structures [20]. Thus, the DNA destabilization energies at protein binding sites of these TFs are expected to be high. This assumption has been verified in yeast by [11] on improving the accuracy of TFBS discovery, which is a different topic other than TF target gene prediction. On the other hand, during transcription, the two DNA strands must be separated to let RNA polymerase slide along the DNA molecule and synthesize a nascent mRNA. Since the binding sites for many general TFs are located in the proximal promoter regions of the transcribed gene, it is expected that the DNA double helix of these regions is low, that is, low DNA duplex stability. Besides, there also exists experimental evidence showing that some regulatory proteins bind to DNA in a single-strand manner [21, 22]. Taken together, these suggest that DNA duplex stability data should be informative of binding sites; whether a lower or higher DNA duplex stability at specific TF binding sites is more preferable depends largely on the binding preference of the TF, that is, whether the TF binds to the the DNA in a double- or single-strand manner. In our study, we assume that TFBSs for TFs with single-strand binding preference occur preferentially in regions with low DNA duplex stability, and the other way around for double-strand binding TFs.

In the TF target gene prediction analysis, the raw DNA duplex destabilization energies were converted into probability values using a Bayesian transformation method, and each TF's binding preference is predicted with a heuristic method (see Section 2 for details).

From the ROC curves shown in Figure 4(a) and supplementary Figure S2(a) we can see that DNA duplex stability alone can slightly improve the TF target gene prediction accuracy, and its performance can be remarkably improved by combining with other priors, such as conservation (Figure 4(c) and supplementary Figure S2(g)) or regulatory potential (Figure 4(b) and supplementary Figure S2(d)). Table 1 also demonstrates that the AUC scores for combining DNA energy with conservation or regulatory potential are higher than those obtained with single additional information sources. These results indicate that DNA duplex stability data has the potential of improving TF target gene prediction depending on how and which data sources it is combined

Table 2: *Sequences used in this study*. One "TFBS duplex stability score" is computed as the average of all the raw DNA duplex stability scores over a given TFBS. The TFBS duplex stability scores are computed for all the binding sites of a promoter sequence. Note that one sequence can have multiple binding TFs and TFBSs, one TF can bind to more than one site, and one TFBS may be recognized by multiple TFs.

Promoter sequence	Length	Binding TFs	TFBS duplex stability scores
AF093878	501	Sp1, Hnf1	5.44, 6.34
Ache	2000	Sp1, Ap2, Egr1	10.03, 9.98, 9.70, 10.03, 10.10, 9.66
Acta1	2000	Srf, Tef, Sp1, Tead1, Sre, Tbp	7.66, 7.71, 7.31, 7.69, 7.71, 7.94, 7.66, 7.77, 7.46, 6.67, 8.09, 7.73, 5.18
Acta2	2000	Carg-d, Prm, Carg-c	9.70, 10.01, 9.76
Actc1	2000	Sp1, Myod1, Srf, Tbp	9.93, 9.82, 9.73, 9.72, 9.23, 9.94, 9.82, 8.90
Alb1	2000	Tcf1, Cebp, Hnf1, Cebp	9.17, 9.80, 9.44, 9.71, 9.64, 9.20
Chrna1	2000	Myf, E1, E2, E3	9.85, 9.91, 9.93, 9.92, 9.92
Chrnb1	2000	Myf, Tef, E1	9.59, 9.78, 9.77, 9.91
Chrnd	2000	Myf, E1	9.90, 9.88, 9.81, 9.76, 9.55, 9.61
Chrne	2000	E1	9.93
Chrng	2000	Myf, E1, E2, E3, E4	9.74, 9.84, 9.64, 9.28, 9.83, 9.86
Ckm	2000	Srf, Nvl, Mef, Prrx1, Myog, Myod1, Myf5, Mef1, Ap2, Myf, Carg3, Mef2-left (-right), E-left (-right), Trp53	9.78, 9.89, 8.15, 8.63, 8.06, 9.94, 9.94, 9.94, 9.95, 9.95, 9.35, 9.94, 9.95, 9.80, 9.97, 9.74, 9.94, 9.69, 8.34, 8.53, 9.80, 9.95, 9.96, 9.87, 9.70
Des	2000	E1, Mef2c, Myod1, Tbp	9.88, 8.23, 6.49, 9.88, 9.88, 8.66
Id2	2000	Cebpb	9.47
Igfbp5	2000	unknown	9.77
M22326	501	Srf	4.69, 4.48, 3.92
M23768	500	Srf, Ap1, Creb	3.99, 5.71, 7.09
M29660	499	Mef2, Caat	5.27, 8.11
M62362	500	Usf, Egr1, Ap2a, Tbp	8.55, 9.60, 9.57, 4.16
M63335	500	Cebp, Nfya, Tbp	2.54, 2.50, 4.34
M86180	514	Sp1	9.97
M86232	428	Srf, Myod1	9.02, 9.87
Mb	2000	Myod1, Mef2, E2, Tbp	8.90, 9.66, 9.73, 8.78, 9.62, 8.34
Myf6	2000	Myf, Myog, Myf5, Myod1	9.62, 9.63, 9.77, 9.63, 9.77, 9.77, 9.63
Myh4	2000	Myf	9.88
Myh6	2000	Mef, Tef, Srf, Mef2, Tead1	7.60, 9.75, 7.60, 8.80, 7.60, 9.75, 8.94
Myl4	2000	E4, E1, Carg	9.97, 9.94, 9.14
Myod1	2000	Ap2, Gc2, Ccaat-box, Sp1, Tbp	9.91, 9.98, 9.55, 9.99, 8.35, 9.55, 9.98
Myog	2000	Myf, Mef, Mef2, E1, Def-2, Myog, Tbp, Myod1	9.03, 7.21, 9.87, 7.00, 9.79, 7.01, 9.79, 9.79, 9.02, 7.00, 7.03, 8.31, 9.79
Q8cfn5	2000	Myf	9.917
Tnnc1	2000	Cef-2, Sp1, Mef2, Mef3, Gata4	9.04, 9.25, 6.54, 9.52, 9.19, 6.54, 9.49, 8.54
Ttr	2000	Cebp, Tcf1, Hnf2, Hnf3, Cebp, Hnf4, Hnf1	9.74, 9.47, 9.50, 9.38, 9.13, 9.74, 9.31, 9.38, 9.50, 9.45, 9.12
U36238	501	Sp1, Cebp, Tbp	10.05, 9.85, 8.78
U69555	505	E2f, Cets	9.92, 9.93
Vim	2000	unknown	9.64, 9.88
X03020	501	Nfkb1, Ap1, Nfat	6.52, 6.63, −0.40, −0.33
X04724	500	Hnf1, Ipf1, Creb, Tbp	5.43, 5.57, 7.27, 8.21
Y18062	500	Hnf1, Cebp	6.38, 0.35, 0.32
NM_010556	500	Oct, Aml, Egr1	2.74, 3.65, 9.77
NM_009715	500	Sp1	9.89, 9.84, 9.81, 9.89, 9.70
NM_008600	500	Sp1, Ap2a, Tbp	9.46, 6.08, 4.55, 3.14, 2.40, 4.10
NM_007398	500	Sp1	10.03, 9.44, 10.02
NM_011358	500	Myb, Caat, Tbp	9.90, 9.74, 6.76, 6.25
NM_023456	500	Sp1, Ap1	10.03, 9.94, 9.92

Table 2: Continued.

Promoter sequence	Length	Binding TFs	TFBS duplex stability scores
NM_011010	500	Olf1	7.41
NM_009415	500	Sp1	5.42, 8.60, 7.87, 9.87
NM_007743	500	Myod1	8.41, 9.10, 8.80, 1.90

Table 3: *Transcription factors used in this study.* "1" and "2" each represents that the corresponding TF binds to DNA in a single and double strand manner, respectively. Empty blank means no literature information is found.

TF	Prediction	Literature	Recognition sequence
AP1	1	1 [22]	GCTCCTCCCA, ATTAATCA, CCCGGGCGTGACTG, TGCGTCA
AP2A	2	2 [23]	GCCGGAGG, CCGCCGGGGTGG, CCCAGGG
CEBPB	2		AATGGCAAT
E2F	2		(CC)TTTCGCGC
EGR1	2	2 [24]	GCGGGGG(CG), TCCCCCCTGCCCCGCCGGGCCCCGCCC
GATA4	2		AGATAG, TGAGATTACA
HNF3	1		AAGTCAATAATC(A), TTTGTGTAGGTTA
IPF1	1		TCTAAT
MEF1	2		CCCCCCAACACCTGCTGCCTGAGCC
MEF2C	1		CTATAAATAC
MYB	2		GAACGT, ACGTTA
MYF5	2	2 [25, 26]	CCCAACACCTGCTGCCTGAGCC, CATCTG, CAGTTG
MYOD1	2	2 [25, 26]	CAACTG, (ATTAACCCA)GACATGTGGC(TGCCCC), CATCTG, (CCCCCCAA)CACCTGCTG(CCTGAGCC), CACTTG, CAGTTG
MYOG	2	2 [25, 26]	(C)CCCAACACCTGCTGCCTGAGCC, CATCTG, CAGTTG
NFAT	1		TTTCCTC
NFKB1	2		GGAGATTCCAC, CCACAACTCA
NFYA	1		CAAT
SP1	2	2 [27]	(T)TCGGGGCGGTGT(G), GCCCCCCAC(CCCTGCCCC), CCGCCC, CCCCACCCCCTGCA, GCGCCAGGGCTGGGCTCCT, CACCTTGGCCACGCCCCTTTGG, CCTGCTTCCCGCCTTTCG, TTTGGTTCCCGCCTCCCCGCCCCC, CCCCTCC(C), TCCTGAAGACCCGCCCTTTTTC, GGCAGAG, CAACC, GGGCGGGGCCGTGGCTCC, GAGCGTGGCGGGCCGCG, (AGGG)TGGGCAG(TCC), GAGGTGGGGGG, AGCCAG, (GGGGGGGGGGGGGGGG)GGGCGG(GGCCGTGGCT), (CCTAAAGTGCTTCCAAA)CTTGGCAAGGGCGAGAGAGGGCGGGTGG
SRF	1		ACCCAAATATGGCT, CCTTACATGG, CCAAGAATGG, CCAAATAAGG, GCCCATGTAAGGAG, GAAACGCCATATAAGGAGCAGG, GCAGCGCCTTATATGGAGTGGC, CTCCAAATTTAGGC, TGCTTCCCATATATGGCCATGT, CCATATTAGG, CTATTATGG
TBP	1		(C)TATAAA(A), TACAAAT, TTAAA, ATAAATA, TTAAAT, TATAAG
TCF1	1		GTTATTGGTTAAAGAAGTATA, GTGTAGGTTACTTATTCTCCTTTTGTTGA
TEAD1	2		(AA)CATTCCTT(CGG), AGGAGGAATGTGC
TRP53	2		GAGCAAGTCA, ATACAAGGCC

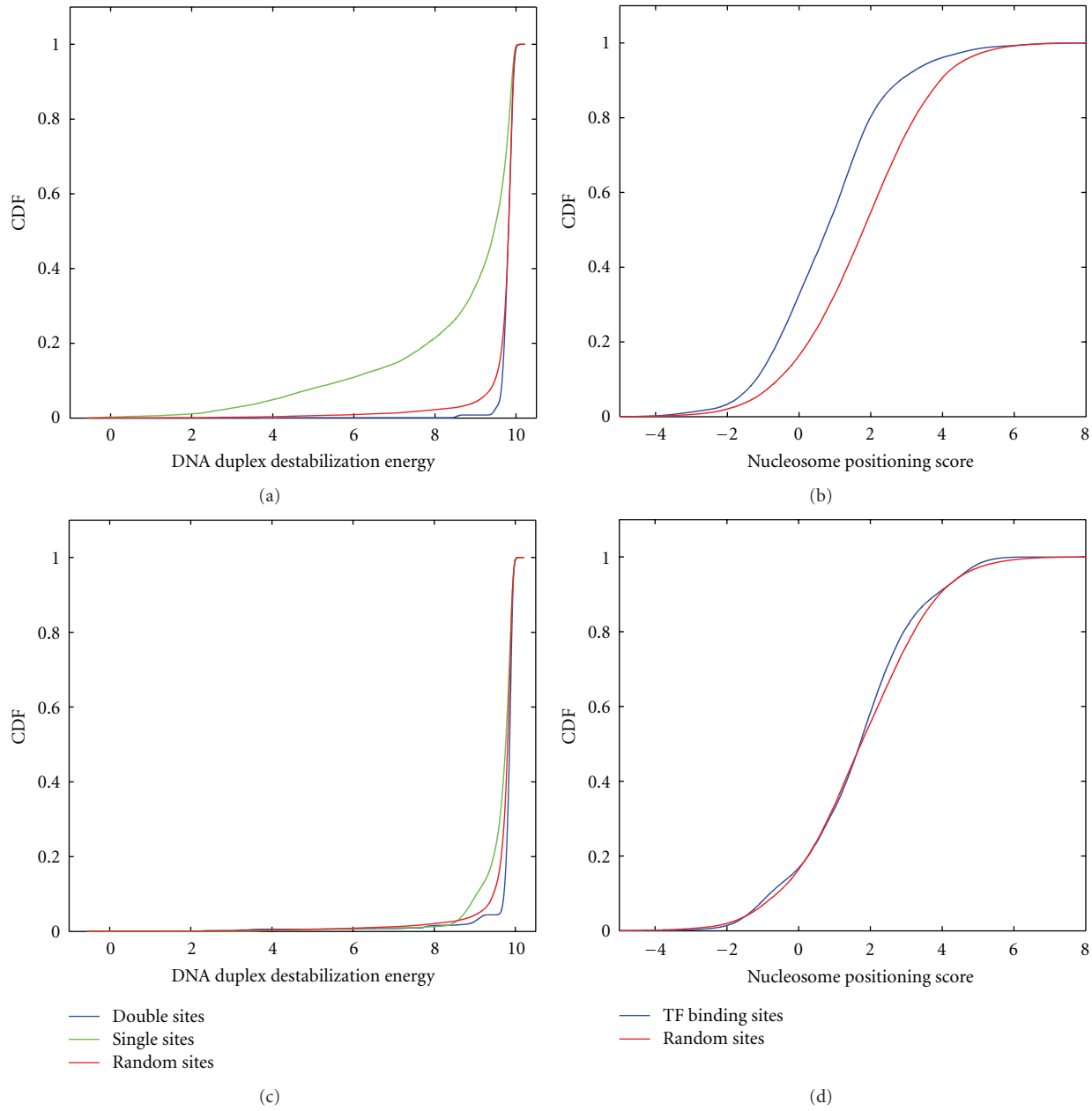

Figure 3: *CDFs of novel information sources at known TFBSs and random sites.* CDFs of (a) DNA duplex destabilization energies at TFBSs of single-strand, double-strand binding TFs, and random DNA sites, (b) nucleosome occupation scores at known TFBSs and random DNA sites. Panels (c) and (d) are similar with (a) and (b), respectively, but with each information scores shifted 100 bps.

with. Further, DNA duplex stabilities are expected to be more informative in TF target gene prediction if they are obtained experimentally.

Out of the 23 TFs whose PSFMs are known and studied here, nine are predicted to bind sequences in a single-strand manner and 14 bind sequences in a double-strand manner. Information such as the names and binding promoters (in mouse genome) of these 23 TFs are listed in Tables 2 and 3, with more detailed information available from http://www.probtf.org/. Also shown in Table 2 are the DNA duplex stability scores for all the binding sites in all the promoter sequences used in this paper, each of which is averaged over all the raw stability scores of a TFBS. These TFs include all the (mouse) TFs whose binding site information can be downloaded from ABS [3] or ORegAnno [5] databases and whose binding specificity model(s) can be found from the TRANSFAC database [4] (professional version 10.2). It is seen from Table 3 that, for the six TFs whose binding preferences are known, our predicted binding preferences accord well with the literature-derived information. In order to avoid the possible bias that could be introduced when the binding preference of each TF is

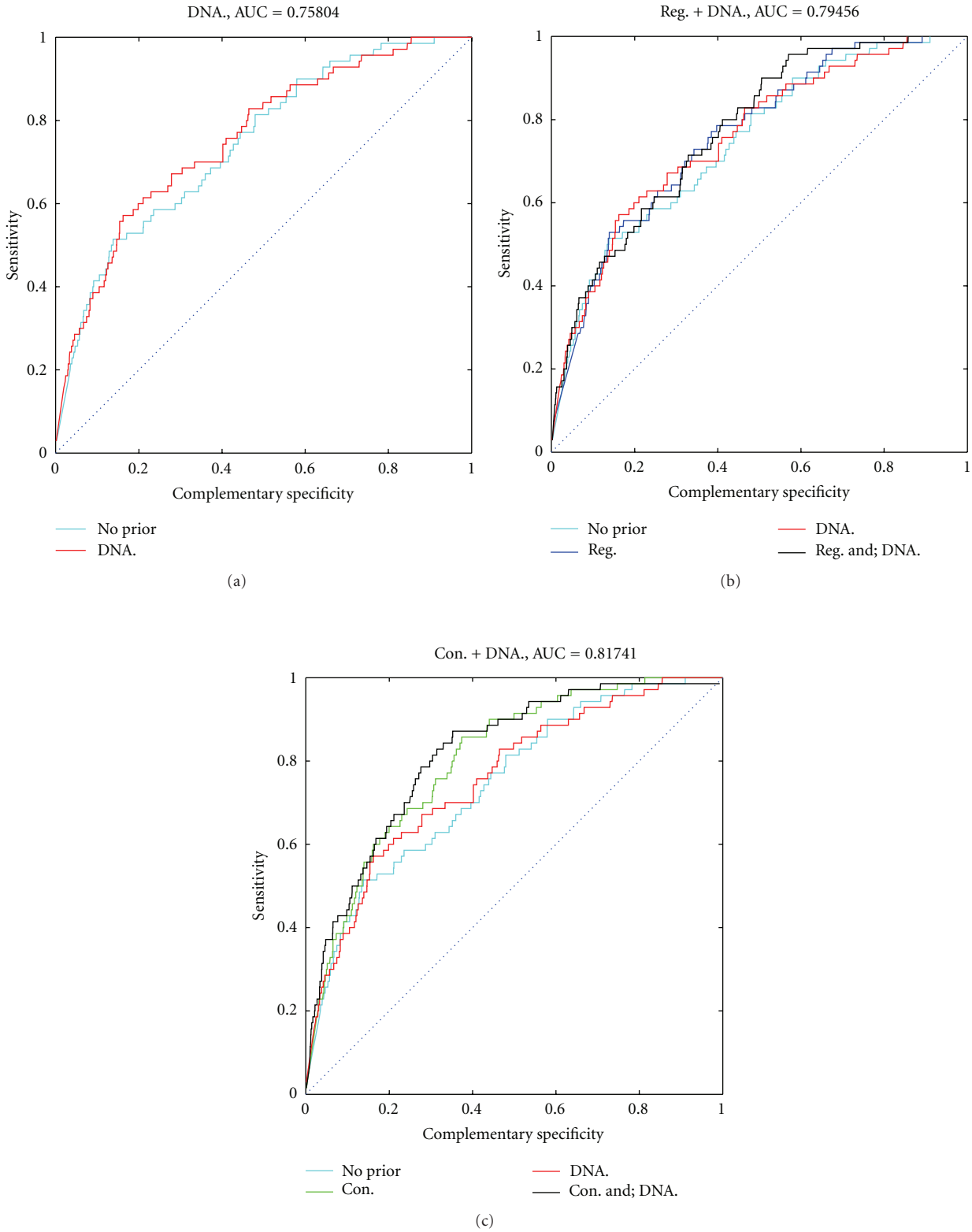

FIGURE 4: *ROC curves for incorporating DNA duplex stability data.* ROC curves of the estimated binding probabilities for the proposed data fusion method when combined with (a) DNA duplex stability data, (b) DNA duplex stability data and regulatory potential, and (c) DNA duplex stability data and evolutionary conservation data.

predicted from the same data that is used for validation, we also performed the standard leave-one-out cross validation on the binding preference prediction. These results clearly demonstrate that no significant differences are observed. Thus, our binding prediction method when integrated with DNA duplex stability data should have a similar good predictive performance outside our test data set as well.

3.1.2. Nucleosome Occupation Prior. Chromatin structure has an important role in regulating the transcriptional machinery. At the genome level, these mechanisms are controlled by the basic structural subunits, nucleosomes, which can limit the access of TFs to their binding sites [1, 17]. Thus, from the viewpoint of computational TFBS prediction, the likelihood of a TF binding to nonfunctional sites can be decreased by locating a stable nucleosome over those genomic regions while keeping functional sites accessible for TF binding. The validity of this assumption can be verified, for example, by the fact that the binding of SP1, GAL4, and USF to nucleosome cores requires other proteins such as nucleoplasmin to remove H2A and H2B which consequently results in nucleosome disassembly [28], and proven by the evidence that the binding propensity of glucocorticoid receptor (GR) to the nucleosome core is much lower than that to the nucleosome free sequence [29]. However, the probability of some TFs binding equally well or even better to sequences occupied by nucleosomes compared with nucleosome free regions could not be excluded, where nucleosome location data alone will not be sufficient and multiple data sources may be used to improve the prediction accuracy.

High-resolution genomewide nucleosome positioning data exist for organisms such as yeast [30] and human [31], but in the case of mouse, we currently need to rely on computational predictions. Indeed, this computational prediction problem has attracted lots of interest and improved methods have been proposed recently. ProbTF method was previously tested with predicted nucleosome locations from Segal's original model which rely on dinucleotide frequencies [1] and the nucleosome data was not found to be informative of binding sites. Here we explore the problem that whether more recent and more accurate nucleosome positioning data together with a novel data fusion method can improve TF target gene prediction. In this study, we used a computational multiresolution method developed in [2] to predict the nucleosome locations for all the 47 tested sequences. We decided to use the raw nucleosome positioning data, that is, without the hidden Markov model (HMM) processing, and employ the extended sequences to obtain the N-score for each genomic location. The raw data were further converted into probabilities using a Bayesian transformation method (for details see Section 2).

We compared the two different nucleosome data by integrating them separately into our TF target gene prediction algorithm. It is particularly promising to see that the use of more accurate nucleosome positioning data from [2] results in more accurate TF target gene prediction as shown in supplementary Figure S3(a). Similarly as in the case of DNA duplex stability data, we combined nucleosome data with conservation (supplementary Figure S3(e)) or regulatory potential data (supplementary Figure S3(c)), and the combined data again improve the TF target gene predictions. For example, the AUC score of 0.7555 which is obtained with nucleosome data alone increases to 0.7946 when combined with regulatory potential, and jumps to 0.8334 when combined with conservation.

In order to gain insight into each individual data source and to assess the extent of possible overfitting problem stemming from parameter optimization, we also prepared an additional control simulation. We shifted each additional data source by 100 base pair positions and then applied our computational methods as explained above, including binding preference prediction and optimization of parameters, to test performance of randomized data. ROC curves corresponding to the four shifted information sources are shown in supplementary Figure S4 and the AUC scores after shifting for each data source are recorded in Table 1. For the two novel data sources, we also compared their CDFs after shifting with the original ones as shown in Figure 3. The Kolmogorove-Smimov statistic (KS statistic) for CDFs of DNA duplex stability scores at random sites and double strand binding sites is 0.3097, and that of random sites and single strand binding sites is 0.3641 (as depicted in Figure 3(a)). However, after shifting, the KS statistics between random locations and double strand binding sites and between random sites and single strand binding sites become 0.1905 and 0.1168, respectively, (see Figure 3(c)). Similarly, the KS statistic between the CDFs of nucleosome positioning data at random sites and nucleosome binding sites is 0.1699 (Figure 3(b)) and drops to 0.0379 after shifting (Figure 3(d)). We also measured the Kullback-Leibler divergence (KL divergence) between each density pair. The KL divergence between PDFs of DNA duplex stability scores at random sites and double and single strand binding sites are 0.1868 and 0.6617 (Figure 3(a)), which decreases to 0.1037 and 0.1065, respectively, after shifting (Figure 3(c)). Likewise, the KL divergence between PDFs of nucleosome positioning data at random sites and nucleosome occupied sites drops from 0.1830 to 0.0330 after shifting, as represented by Figures 3(b) and 3(d), respectively. These results show that no information is gained from the shifted data sources. Taken together with the cross-validation results shown above, this demonstrates that the improved binding prediction accuracy is not an artifact of overfitting.

We further compared the scaling parameter a (see (11) in Section 2) when integrating different nucleosome data and DNA duplex stability data into the TF target gene prediction framework. The parameter a essentially determines the weight of each individual information source. As shown in Table 1, parameter a of nucleosome positioning data obtained from [2] (0.04) is higher than that obtained with data from [1] (0.01), which is consistent with results in supplementary Figure S3(a) where nucleosome data from [2] clearly provides more information than those obtained from [1]. Similarly, parameter a of DNA duplex stability data for TFs with single-strand binding pattern (0.06) is higher than that for TFs with double-strand binding pattern (0.01). This

is again consistent with results shown in Figure 3(a), where DNA stability energies of single-strand binding TFs provide much better discrimination than those of double-strand binding TFs. These results show that the scaling parameter a has an association with data quality, where a higher a indicates a more informative data.

3.2. Multiple Data Fusion Method. We next briefly demonstrate the performance of the new data fusion method and compare it with that of a standard weighting-based scheme proposed in [12]. Qualitatively, the previous data fusion method is based on a type of averaging where a genomic location is suggested to contain a binding site only if a large majority of the additional data sources indicate a binding site, whereas the new method can assign more prior probability to a genomic location if it is indicated as a binding site by a few (or even a single) more informative data sources (see Section 2 for a detailed technical description of our data fusion methods).

The performance of the old and new data fusion methods are illustrated in supplementary Figure S1, which shows the ROC curves for finding the verified binding sites in the gene promoters set using both evolutionary conservation and regulatory potential. Parameters in supplementary Figures S1(a) and S1(c) are chosen by the whole AUC and the AUC30, respectively. Supplementary Figure S1(a) shows that the new method works better than the old one by generating higher overall AUC, and supplementary Figure S1(c) demonstrates that the new method can improve the prediction accuracy especially in low false positive rate (FPR) region, which is a highly preferable property in general.

Supplementary Figures S1(b) and S1(d) show the histograms of the predicted binding probabilities for both the old and new data fusion methods, where the parameters in Figures S1(b) and S1(d) are selected according to the whole AUC and AUC30, respectively. Histograms are drawn separately for negative and positive cases and, hence, these graphs clearly demonstrate how well the two methods are able to discriminate the target genes that contain known binding sites from nontarget genes that do not contain binding sites. From these graphs, we can see that the new method improves discrimination by assigning much smaller binding probabilities for sequences with no known binding sites (no matter whether AUC or AUC30 is used), which thus results in much smaller false positive rate. AUC scores for single and all combinations of multiple data sources are summarized in an ascending order in Table 1, and their corresponding data fusion results are shown and discussed in the following sections.

3.3. Comparison of Combinations of Information Sources. In order to better understand that which combinations of additional genome-level data sources are most informative of TFBSs, we compared the TF target gene prediction accuracy of all possible combinations among evolutionary conservation, regulatory potential, nucleosome locations, and DNA duplex stability. The best combination is conservation and nucleosome positioning, whose results have already been shown in supplementary Figure S3(e).

Results for all the six duplets of data sources are reported in supplementary Figure S5, which shows that most of the combinations of two data sources work better than their corresponding single data sources except for the combination of nucleosome occupation and DNA energy. This suggests that certain redundancy might exist between nucleosome occupation and DNA energy, which is not entirely surprising since a DNA region that is not within a nucleosome is likely to need less energy to destabilize the two strands than DNA within a nucleosome. This motivates us to group the four information sources into two categories, where group 1 includes evolutionary conservation and regulatory potential, and group 2 includes nucleosome locations and DNA duplex stability. Our results indicate that when a pair of data sources come from different groups, that is, have little redundancy, their joint performance can be better than those of their corresponding single data sources. Moreover, the best performance is achieved with a pair of additional data sources (supplementary Figure S5(b)), and adding more information sources into this pair cannot further improve the accuracy. The above results and analysis suggest that combining data sources that are redundant does not necessarily improve the overall performance. In other words, in order to gain a better prediction accuracy it is better to combine data sources that provide information from different perspectives of the same biological system.

Results for all four triplets of data sources are shown in supplementary Figure S6, which all perform better than their corresponding single data sources. It is seen that the best result is obtained by combining conservation, regulatory potential and nucleosome positioning, which accords well with our expectation since "conservation and regulatory potential" is the most informative pair in the lower false positive region (supplementary Figure S5(f)), and "nucleosome positioning, and regulatory potential" forms the best pair with respect to higher false positive region (supplementary Figure S5(d)).

Supplementary Figure S7 shows the ROC curve for the only quartet. Although one could expect that adding more information sources into TF target gene prediction always improves the prediction accuracy, our results show that it is not always the case. This finding is understood by realizing the difficulty of combining complex and poorly characterized genome-level data sources into TF target gene prediction.

4. Conclusions

We have three main contributions in this paper. Firstly, we have developed a new data integration method for TF target gene prediction from multiple data sources. The new method is compared with the one employed in [12] using a TF target gene prediction algorithm called ProbTF [12], and the results show that the new data fusion principle improves the previous method by lower false positive rate. Secondly, we have demonstrated the use of two novel information

sources, DNA duplex stability and raw nucleosome occupancy predictions from a method proposed in [2], to guide TF target gene predictions. Our results show that both nucleosome occupancy and DNA stability data can improve TF target gene prediction accuracy especially when combined with evolutionary conservation or with conservation and regulatory potential. Moreover, more accurate nucleosome predictions result in better TF target gene predictions. It is also worth noticing that we do not distinguish different TFs regarding data source usage except for DNA duplex stabilities, where double or single strand binding proteins are treated differently and a heuristic method is adopted to classify them. Thirdly, we have compared all the possible combinations among conservation, regulatory potential, nucleosome positioning and DNA stability, whose results can be availed in data source selection or preparation when dealing with data integration problem in a particular application. We grouped the four tested information sources into two categories based on biological arguments: group 1 contains conservation and regulatory potential, and group 2 consists of nucleosome locations and DNA duplex stability. We found that combining data from different groups is more likely to improve TF target gene predictions presumably because data sources between the two groups are less redundant.

Although the assumption that all TFs bind to DNA in double-strand manner works well in yeast [11], it may not be sufficient in higher organisms, such as mouse, as shown in this study (see, e.g., Figure 3(a)). Instead, we obtained informative DNA duplex stability prior by assuming different binding preferences for different TFs. We constructed the binding preference of each TF with a simple heuristic which assesses the binding preference for a TF from a set of known binding sites. We have used cross-validation and an additional base pair shifting simulations to show that binding preference prediction and parameter optimization do not result in any (optimistic) bias, or overfitting, in binding prediction accuracy. However, the use of the DNA duplex stability data is limited because little verified information about TF binding specificities can be found from the literature and, therefore, binding specificities need to be learned from the data as well which currently requires that a set of verified binding sites is known. Future research goals include to develop an (unsupervised) algorithm for predicting the binding preference for TFs without prior knowledge of the known binding sites. Moreover, it is possible that one TF may have multiple folding modes, and can bind different sequences with different patterns. For example, MyoD, a member of helix-loop-helix protein family, can not only recognize the double-stranded DNA-binding site (called E-box) in many muscle and nonmuscle genes, but also bind to the noncoding strand of an E-box from the muscle-specific creatine kinase enhancer in a single-stranded manner [25]. To take this possibility into account, a more sophisticated assumption can be applied; that is, TFs can have different binding preferences to different sequences or under different experimental conditions. In this direction, we can also try to incorporate other data sources, such as ChIP-chip data, into our data fusion framework.

Nucleosome positioning data is employed in this study assuming that nucleosomes compete with DNA binding proteins [1] for target DNA binding sites. Although this assumption is generally true, we could not exclude the possibility that some TFs may selectively bind to nucleosome-occupied regions. Binding sites of such TFs, if exist, can not be recognized by the method presented here when employing nucleosome occupancy data, but can be rescued, for example, by incorporating other information sources.

Authors Contributions

X. Dai and H. Lähdesmäki designed the study and prepared the paper. O. Yli-Harja participated in the study design. X. Dai developed the new data fusion method, implemented the two novel data sources in TF target gene prediction, and performed all the simulations.

Acknowledgments

The authors would like to thank Yuan Guo-Cheng for providing us his software for nucleosome occupation prediction. This work was supported by Tampere Graduate School in Information Science and Engineering (TISE) (XFD) and the Academy of Finland (Grant no. 213462).

References

[1] Y. Hayashi, N. Sano, and M. Horikoshi, "A genomic code for nucleosome positioning," *Chemtracts*, vol. 19, no. 6, pp. 223–233, 2007.

[2] G. C. Yuan and J. S. Liu, "Genomic sequence is highly predictive of local nucleosome depletion," *PLoS Computational Biology*, vol. 4, no. 1, article e13, 2008.

[3] 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 Research*, vol. 34, pp. D63–D67, 2006.

[4] E. Wingender, X. Chen, R. Hehl et al., "TRANSFAC: an integrated system for gene expression regulation," *Nucleic Acids Research*, vol. 28, no. 1, pp. 316–319, 2000.

[5] S. B. Montgomery, O. L. Griffith, M. C. Sleumer et al., "ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation," *Bioinformatics*, vol. 22, no. 5, pp. 637–640, 2006.

[6] K. D. MacIsaac and E. Fraenkel, "Practical strategies for discovering regulatory DNA sequence motifs.," *PLoS Computational Biology*, vol. 2, no. 4, p. e36, 2006.

[7] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep III, and M. L. Bulyk, "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nature Biotechnology*, vol. 24, no. 11, pp. 1429–1435, 2006.

[8] C. T. Harbison, D. B. Gordon, T. I. Lee et al., "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 430, no. 7004, pp. 99–104, 2004.

[9] Y. Qi, A. Rolfe, K. D. MacIsaac et al., "High-resolution computational models of genome binding events," *Nature Biotechnology*, vol. 24, no. 8, pp. 963–970, 2006.

[10] L. Narlikar, R. Gordân, and A. J. Hartemink, "A nucleosome-guided map of transcription factor binding sites in yeast.," *PLoS Computational Biology*, vol. 3, no. 11, p. e215, 2007.

[11] R. Gordân and A. J. Hartemink, "Using DNA duplex stability information for transcription factor binding site discovery," in *Proceedings of Pacific Symposium on Biocomputing (PSB '08)*, pp. 453–464, World Scientiffic, 2008.

[12] H. Lähdesmäki, A. G. Rust, and I. Shmulevich, "Probabilistic inference of transcription factor binding from multiple data sources," *PLoS One*, vol. 3, no. 3, article e1820, 2008.

[13] A. Siepel, G. Bejerano, J. S. Pedersen et al., "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes," *Genome Research*, vol. 15, no. 8, pp. 1034–1050, 2005.

[14] J. Taylor, S. Tyekucheva, D. C. King, R. C. Hardison, W. Miller, and F. Chiaromonte, "ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements," *Genome Research*, vol. 16, no. 12, pp. 1596–1604, 2006.

[15] C. J. Benham and C. Bi, "The analysis of stress-induced duplex destabilization in long genomic DNA sequences," *Journal of Computational Biology*, vol. 11, no. 4, pp. 519–543, 2004.

[16] C. Bi and C. J. Benham, "WebSIDD: server for predicting stress-induced duplex destablized (SIDD) sites in superhelical DNA," *Bioinformatics*, vol. 20, no. 9, pp. 1477–1479, 2004.

[17] C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb, "Evidence for nucleosome depletion at active regulatory regions genome-wide," *Nature Genetics*, vol. 36, no. 8, pp. 900–905, 2004.

[18] W. W. Wasserman and A. Sandelin, "Applied bioinformatics for the identification of regulatory elements," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 276–287, 2004.

[19] D. L. Ollis and S. W. White, "Structural basis of protein-nucleic acid interactions," *Chemical Reviews*, vol. 87, no. 5, pp. 981–995, 1987.

[20] G. Wisedchaisri, R. K. Holmes, and W. G. J. Hol, "Crystal structure of an IdeR-DNA complex reveals a conformational change in activated IdeR for base-specific interactions," *Journal of Molecular Biology*, vol. 342, no. 4, pp. 1155–1169, 2004.

[21] R. Duncan, L. Bazar, G. Michelotti et al., "A sequence-specific, single-strand binding protein activates the far upstream element of c-myc and defines a new DNA-binding motif," *Genes and Development*, vol. 8, no. 4, pp. 465–480, 1994.

[22] L. M. E. Finocchiaro, P. Amati, and G. C. Glikin, "Single strand binding protein specific for the polyoma early-coding strand of PEA1 (AP1) regulatory sequence," *Nucleic Acids Research*, vol. 19, no. 15, pp. 4279–4287, 1991.

[23] A. B. Heimberger, E. C. McGary, D. Suki et al., "Loss of the AP-2α transcription factor is associated with the grade of human gliomas," *Clinical Cancer Research*, vol. 11, no. 1, pp. 267–272, 2005.

[24] B. Christy and D. Nathans, "DNA binding site of the growth factor-inducible protein Zif268," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 22, pp. 8737–8741, 1989.

[25] K. Walsh and A. Gualberto, "MyoD binds to the guanine tetrad nucleic acid structure," *Journal of Biological Chemistry*, vol. 267, no. 19, pp. 13714–13718, 1992.

[26] L. A. Sabourin and M. A. Rudnicki, "The molecular regulation of myogenesis," *Clinical Genetics*, vol. 57, no. 1, pp. 16–25, 2000.

[27] K. J. Perkins, E. A. Burton, and K. E. Davies, "The role of basal and myogenic factors in the transciptional activation of utrophin promoter A: implications for therapeutic up-regulation in Duchenne muscular dystrophy," *Nucleic Acids Research*, vol. 29, no. 23, pp. 4843–4850, 2001.

[28] H. Chen, B. Li, and J. L. Workman, "A histone-binding protein, nucleoplasmin, stimulates transcription factor binding to nucleosomes and factor-induced nucleosome disassembly," *EMBO Journal*, vol. 13, no. 2, pp. 380–390, 1994.

[29] Q. Li and O. Wrange, "Translational positioning of a nucleosomal glucocorticoid response element modulates glucocorticoid receptor affinity," *Genes and Development*, vol. 7, no. 12A, pp. 2471–2482, 1993.

[30] W. Lee, D. Tillo, N. Bray et al., "A high-resolution atlas of nucleosome occupancy in yeast," *Nature Genetics*, vol. 39, no. 10, pp. 1235–1244, 2007.

[31] D. E. Schones, K. Cui, S. Cuddapah et al., "Dynamic regulation of nucleosome positioning in the human genome," *Cell*, vol. 132, no. 5, pp. 887–898, 2008.