BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 5 2012, pages 701–708
doi:10.1093/bioinformatics/bts002

Systems biology

Advance Access publication January 11, 2012

De novo motif discovery facilitates identification of interactions between transcription factors in *Saccharomyces cerevisiae*

Mei-Ju May Chen[1], Lih-Ching Chou[2], Tsung-Ting Hsieh[3], Ding-Dar Lee[2], Kai-Wei Liu[2], Chi-Yuan Yu[4], Yen-Jen Oyang[1,2,4,5], Huai-Kuang Tsai[6,*] and Chien-Yu Chen[1,3,5,*]

[1]Genome and Systems Biology Degree Program, [2]Department of Computer Science and Information Engineering, [3]Department of Bio-Industrial Mechatronics Engineering, [4]Graduate Institute of Biomedical Electronics and Bioinformatics, [5]Center for Systems Biology, National Taiwan University and [6]Institute of Information Science, Academia Sinica, Taipei, Taiwan

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Gene regulation involves complicated mechanisms such as cooperativity between a set of transcription factors (TFs). Previous studies have used target genes shared by two TFs as a clue to infer TF–TF interactions. However, this task remains challenging because the target genes with low binding affinity are frequently omitted by experimental data, especially when a single strict threshold is employed. This article aims at improving the accuracy of inferring TF–TF interactions by incorporating motif discovery as a fundamental step when detecting overlapping targets of TFs based on ChIP-chip data.

Results: The proposed method, simTFBS, outperforms three naïve methods that adopt fixed thresholds when inferring TF–TF interactions based on ChIP-chip data. In addition, simTFBS is compared with two advanced methods and demonstrates its advantages in predicting TF–TF interactions. By comparing simTFBS with predictions based on the set of available annotated yeast TF binding motifs, we demonstrate that the good performance of simTFBS is indeed coming from the additional motifs found by the proposed procedures.

Contact: hktsai@iis.sinica.edu.tw; chienyuchen@ntu.edu.tw

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 26, 2011; revised on December 31, 2011; accepted on January 02, 2012

1 INTRODUCTION

Transcriptional regulation plays a key role in gene expression. It typically takes place after the binding of transcription factors (TFs) to specific promoter regions of genes. Usually, gene expression is regulated by a group of TFs instead of a single TF. When the influence of one TF on its target genes depends on the presence or absence of another TF, it is said that these two TFs have interactions in gene regulation (Wang *et al.*, 2009). This article aims at discovering TF–TF interactions that happen on the TF binding sites (TFBSs) of the target genes the two TFs share. Correct predictions of such interactions are important in studying gene regulation.

So far, several types of TF–TF interactions have been observed and discussed. As exemplified in Figure 1, two TFs, TF1 and TF2, might form a complex and then bind to a common TFBS, which is often represented by a TF binding motif (Fig. 1a). Alternatively, they might utilize the DNA-binding domain of TF1 to bind to DNA (Fig. 1b). In this case, we say that TF2 indirectly binds to the TFBS (piggy back) (Kar and Adhya, 2001). In both Figure 1a and Figure 1b, the two TFs interact with each other through direct protein–protein interactions (PPIs). On the other hand, it is possible that both TF1 and TF2 have their own DNA-binding domains and respective TF binding motifs. As exemplified in Figure 1c, TF1 and TF2 may not have direct contact but transcription is affected when either one is absent. In this type of TF–TF interaction, the TFs regulate expression of the target genes without directly interacting with each other. Figure 1d demonstrates the situation that TF1 and TF2 compete for the same TFBS. Furthermore, more complicated situations might exist. For example, in Figure 1e, TF1 and TF2 alternatively cooperate with another TF to regulate genes.

Correct prediction of TF–TF interactions is a prerequisite for understanding transcriptional regulation. According to the fact that an interacting TF pair would jointly regulate a set of target genes, it is intuitively considered that the interaction of a TF pair can be detected by exploring the degree of shared target genes (Datta and Zhao, 2008). Before going into interaction inference, a reliable list of the target genes of a TF should be constructed. Recent advances of high-throughput tools provide such valuable information. For example, chromatin immunoprecipitation chip (ChIP-chip) measures the binding affinity of TFs on specific DNA sequences *in vivo* (Buck and Lieb, 2004; Harbison *et al.*, 2004; Ren *et al.*, 2000). Two groups previously published considerably large datasets of ChIP-chip for yeast (Harbison *et al.*, 2004; Lee *et al.*, 2002). These resources have been widely utilized to predict TF binding motifs (Chen *et al.*, 2008; Eden *et al.*, 2007; MacIsaac *et al.*, 2006; Narlikar *et al.*, 2006; Linhart *et al.*, 2008) and discover cooperating TFs (Balaji *et al.*, 2006; Banerjee and Zhang, 2003; Chang *et al.*, 2006; Datta and Zhao, 2008; Kato *et al.*, 2004; Pilpel *et al.*, 2001; Tsai *et al.*, 2005) in recent years. Another useful technique to effectively detect TF–DNA interactions *in vitro* is protein binding microarrays (PBMs) (Berger *et al.*, 2006; Bulyk, 2007; Bulyk *et al.*, 1999, 2001; Mukherjee *et al.*, 2004). Badis *et al.* (2008) and Zhu *et al.* (2009), respectively, produced datasets of PBM experiments and discovered yeast TFBSs based on the detected DNA sequences.

*To whom correspondence should be addressed.

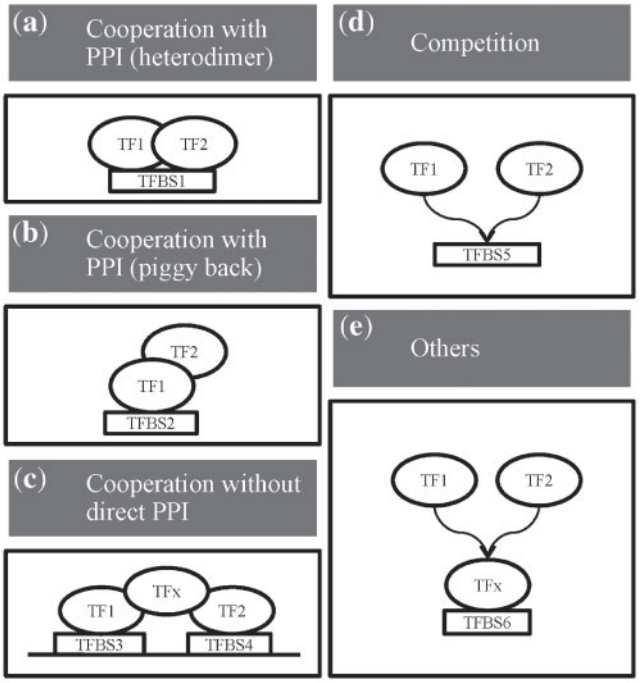

Fig. 1. Transcription could be regulated by multiple TFs through direct or indirect interactions. (**a**) TF1 and TF2 form complexes first and bind to DNA through a common TF binding site (TFBS1); (**b**) the complex of TF1 and TF2 utilizes the DNA-binding domain of TF1 to bind to TFBS2. It is said that TF2 indirectly binds to TFBS2; (**c**) TF1 binds to its own binding site (TFBS3), and TF2 binds to its own binding site (TFBS4) on the same promoter. These two TFs do not directly bind to each other; (**d**) TF1 and TF2 compete for TFBS5; and (**e**) a more complicated example: TF1 and TF2 alternatively cooperate with another TF to regulate genes.

No matter which technology is adopted, one of the main issues involved in the prediction of TF–TF interactions has been how to collect target genes of a TF accurately. When handling ChIP-chip data, many studies apply a consistent cutoff *P*-value to assign genes as the targets of a TF (Chang *et al.*, 2006; Lemmens *et al.*, 2006; McCord *et al.*, 2007; Tsai *et al.*, 2005; Wu *et al.*, 2007). However, this strategy might miss true target genes (those with low binding affinity in ChIP-chip data) if the threshold is strict, or include fake target genes in the list if the threshold is loose (Datta and Zhao, 2008). Accordingly, Datta *et al.* proposed a method to dynamically determine the threshold for collecting target gene set for a TF based on the Expectation–Maximization (EM) algorithm (Datta and Zhao, 2008). Nevertheless, the performance of employing dynamic thresholds for determining target genes to infer TF cooperation has not been comprehensively evaluated against literature.

Since TFBSs underlie the links between a TF and its target genes, this article aims at incorporating automated motif discovery when inferring TF–TF interactions. In other words, to more accurately predict TF–TF interactions, we first checked if a motif is respectively enriched within the potential target genes of both of the two TFs and considered the common motif that intuitively support the predicted interaction. The proposed method, simTFBS, employs motif discovery during ChIP-chip data analysis to improve the quality of target gene lists identified. With the automatically discovered motifs, simTFBS recruits more genes with low binding affinity from ChIP-chip data by requiring that the promoter region of the gene must contain the *de novo* identified motif. In the end, each predicted TF–TF interaction is reported along with the potentially involved TF binding motifs, more than half of which were observed to be similar to known regulatory motifs. The analyses conducted in this study show that we can identify more known interacting TF pairs without sacrificing the specificity of the methodology.

2 MATERIALS AND METHODS

2.1 Dataset

ChIP-chip provides the information of genome-wide binding locations for a given protein *in vivo*. A potential list of target genes of a TF can be collected from ChIP-chip data by applying a fixed threshold on the binding significance (*P*-values) (Chang *et al.*, 2006; Lemmens *et al.*, 2006; McCord *et al.*, 2007; Tsai *et al.*, 2005; Wu *et al.*, 2007). The examples shown in Figure 1 imply that interacting TFs might be discovered by investigating the intersection of the two target gene sets derived from ChIP-chip data. In this study, we employed the ChIP-chip data from the study of Harbison *et al.* (2004). This dataset provides DNA-binding probabilities of 203 TFs under rich medium (YPD) and 84 experiments under other conditions in yeast. In this study, the ChIP-chip data of 203 TFs under all types of conditions (350 chips) were used.

2.2 The proposed method: simTFBS

First, simTFBS collected lists of potential target genes based on ChIP-chip data and generated putative binding motifs by motif discovery. With the set of discovered potential motifs of each TF, we identified interacting TF pairs based on the assumption that interacting TFs would share a set of common genes and thus the motif discovery method might discover similar TF binding motifs. The flow chart of simTFBS is shown in Figure 2, and the detailed procedures are illustrated as follows.

Motif discovery based on gene sets collected from ChIP-chip (Procedure 1 in Fig. 2): for each TF, genes with $P < 0.001$ (or > 0.9) in its ChIP-chip data were collected as the positive (or negative) set. It was required that the number of genes in any positive set is >30. The setting of 30 genes was desirable to fulfill the requirement of sufficient positive sequences when conducting motif discovery. It was observed that 30 is close to the minimum number of positive sequences required to deliver satisfactory results of motif discovery. If this condition was not fulfilled, the *P*-value cutoff (Pc) was adjusted dynamically until it is satisfied. For all the genes in both of the positive and negative sets, the 500 bp upstream sequences were considered as the promoters and collected. Next, we executed motif discovery by the algorithm proposed by Chen *et al.* (2008) to generate 10 top-scored motifs for a TF. This algorithm (eTFBS) was adopted because it has been shown to achieve high accuracy in identifying TF-binding motifs from ChIP-chip data. eTFBS uses a novel hybrid ranking system that considers not only the preferential occurrence of a motif in a set of target sequences over a set of non-target sequences, but also the binding strength of a TF to a putative target promoter and the degree of evolutionary conservation of a predicted motif. Readers can refer to Chen *et al.* (2008) for detailed parameter settings. Here, for each TF, a list of 10 top-scored motifs with length longer than six was reported. The predicted motifs were enriched in the positive sequences (the promoters of high-affinity genes). The definition of enrichment was set as 25%, i.e. 25% of positive sequences must contain the discovered motifs. These motifs are then used for the next procedure. In eTFBS, the specific positions are grouped into two short motifs interleaved with a long gap. In this study, the maximum gap length was set as 15. A motif in eTFBS is represented as a collection of compatible consensus sequences. Substrings in promoter sequences were detected as predicted TFBSs (also called motif instances) if a match was found to consensus sequences (e.g. GCGnATC) using regular expressions. In this study, a motif is also represented as a

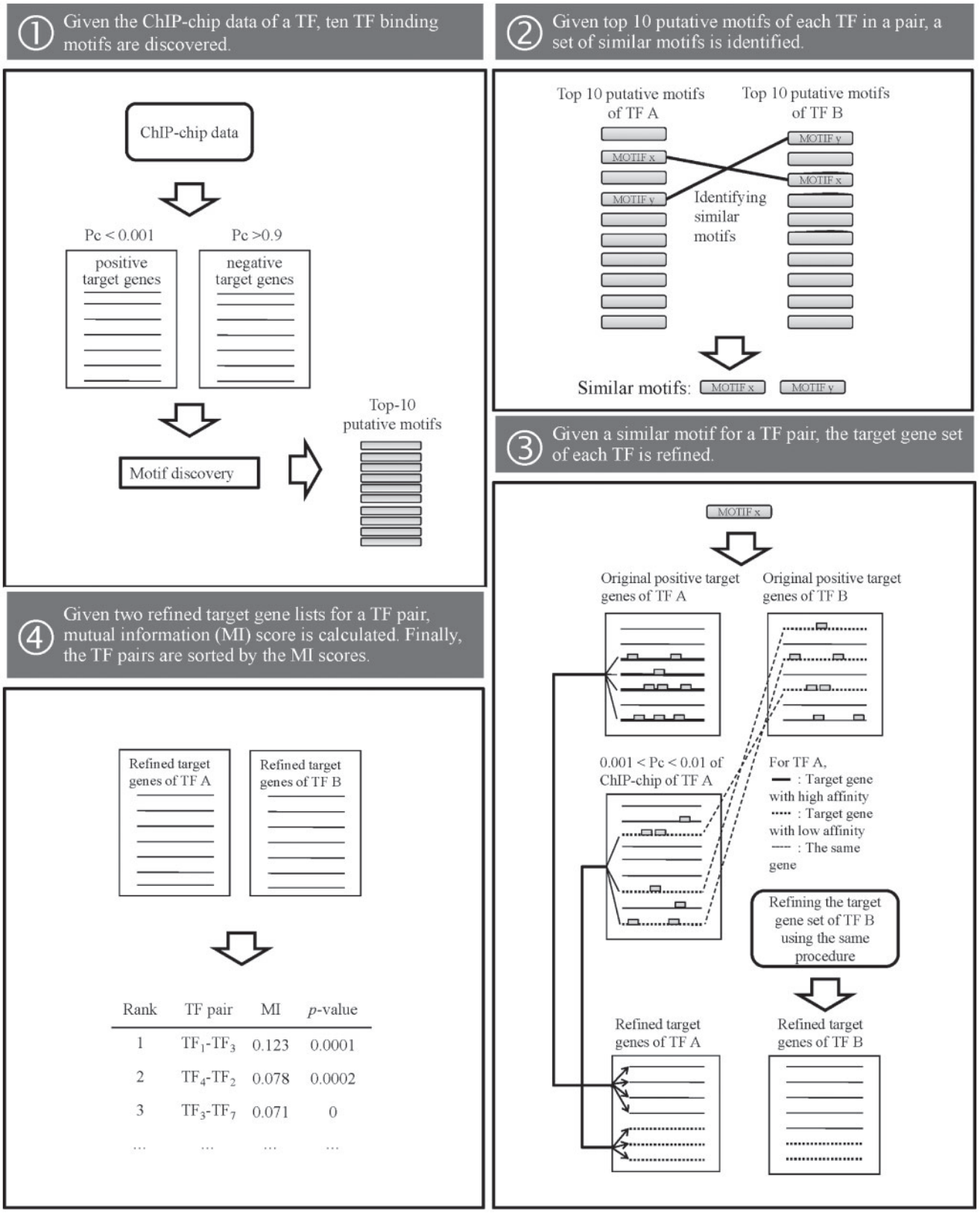

Fig. 2. The flow chart of the proposed method, simTFBS.

position frequency matrix (PFM), which was constructed by the set of predicted TFBSs from different target genes.

Identification of similar TF binding motifs for a TF pair (Procedure 2 in Fig. 2): to be a pair of TFs for interaction prediction, it was required in advance that the number of common target genes from the respective positive sets collected in 'procedure 1' must be greater than zero. After that, we checked whether there exists a pair of similar motifs across the motif lists of the two TFs. The similarity of two motifs was examined by the following two scores: (i) the similarity score between their PFMs; and (ii) the degree of overlap between two predicted TFBS lists within the promoters of the common genes. For a TF pair, we first calculated the similarity scores of PFMs, as in Chen *et al.* (2008), between any pair of the 10 top-scored motifs from the two TFs. The TF pairs with at least one score (100 combinations in total) >0.8 are preserved. Then, for each motif pair with the similarity score >0.8, the predicted TFBSs in promoter regions of the common genes were respectively collected. A common TFBS was claimed as long as the motif instances from two TFs shared at least half of the positions (nucleotides). Let N_A (N_B) be the number of TFBSs from TF A (TF B), and N_C be the number of common TFBSs. It is supposed that the two lists of motif instances would be highly concurrent if the two motifs represent the same TF binding motif. In this regard, two motifs are considered as the same if the overlap score, $[(N_C/N_A) + (N_C/N_B)]/2$, is greater than a preset threshold, which was empirically set as 0.7 in this study. It is possible that one motif from TF A is similar to more than one motif from TF B. If it happens, all the similar motifs are merged into one PFM. In the end of Procedure 2, we reported a list of non-redundant TF binding motifs for each of the potential interacting TF pairs.

Refinement of target gene lists by the discovered common motifs for a TF pair (Procedure 3 in Fig. 2): the set of common TF binding motifs were utilized to refine the target gene set of each TF in a pair. For TF A (or TF B), we used one of the common motifs (e.g. MOTIFx) to scan the promoter regions of the genes obtained from the ChIP-chip data of TF A (or TF B). A gene in the original positive target gene set of TF A (or TF B), called high-affinity genes of TF A (or TF B), would be selected in the refined list of TF A (or TF B) if it contains MOTIFx on its promoter region. Meanwhile, a Pc-hybrid procedure is designed to identify low-affinity genes as follows: for a pair of TFs (TF A and TF B), a gene is a low-affinity gene of TF A (TF B) if it has a *P*-value ranging from 0.001 to 0.01 in the ChIP-chip data of TF A (TF B) and is a high-affinity gene of the other TF in the pair, TF B (TF A). A low-affinity gene of TF A (or TF B) will be preserved in the refined list of TF A (or TF B) if it contains MOTIFx on its promoter region.

TF-TF interaction prediction based on refined target gene lists (Procedure 4 in Fig. 2): based on the refined target sets of each TF in a pair, we calculated the mutual information (MI) of each TF pair using the following equation, provided in the study of Manke *et al.* (2003). In this equation, $I(x, y)$ denotes the MI for a TF pair (x, y),

$$I(x,y)=\sum_{\alpha,\beta} p_{\alpha\beta}(x,y)\left(\log_2 \frac{p_{\alpha\beta}(x,y)}{p_\alpha(x)p_\beta(y)}\right) \quad (1)$$

where α and β, indicate the binding information of TF x and TF y, respectively, and take values in the Boolean alphabet (0, 1); $p(x)$ represents the frequency distribution of binding information (0 or 1) for TF x; and $p(x, y)$ is the distribution over all possible combinations (00, 01, 10, or 11) for a pair (x, y). Furthermore, a permutation test was constructed for estimating the *P*-value of MI scores. For each TF pair, the background distribution is constructed by 10,000 simulated MI scores, which were calculated by utilizing randomly chosen target genes in an equal size of the two target gene lists for both TFs in the pair. TF pairs with $P<0.01$ would be considered as interacting TF pairs. It is possible that a TF pair might have more than one similar motif pair to support their interaction. In such cases, each similar motif pair will go through Procedures 3 and 4 to calculate the MI score. Finally, the largest MI score is taken for the query TF pair. Furthermore, the predictions for a TF pair in different conditions are combined into a single prediction. In this regard, the highest MI score is taken. In addition, TF–TF interactions with MI scores >0.005 were categorized as predictions with high confidence. Finally, the predicted TF pairs are sorted based on their MI scores instead of *P*-values in descending order, since all the *P*-values of the 221 high-confident interacting pairs were highly significant with values close to zero.

2.3 Naïve methods for comparison

For evaluating the performance of the proposed method, we considered three naïve methods of determining the sets of potential target genes for comparison. Pc-0.001 and Pc-0.01 use a simple cutoff of *P*-values from the ChIP-chip data as the threshold for identifying target genes (Pc-0.001 means $P<0.001$ and Pc-0.01 means $P<0.01$). The third method, Pc-hybrid, is a hybrid approach that conditionally includes genes with lower affinity as the target genes of a TF. For a pair of TFs (TF A and TF B), a gene is considered as a low-affinity gene of TF A (TF B) if it has a *P*-value ranged from 0.001 to 0.01 in the ChIP-chip data of TF A (TF B) and is a high-affinity ($P < 0.001$) gene of TF B (TF A).

2.4 Collecting TF–TF interactions with literature support for performance evaluation

To validate the proposed method, we collected the annotated interacting TF pairs from heterogeneous sources and constructed the validation lists of TF–TF interactions which mainly consist of protein–protein interacting pairs, synergistic TF pairs and piggy-back TF pairs. Details of the collection of validation lists are described in the Supplementary Material.

To be a TF pair in the validation list, we required that both TFs must have ChIP-chip data from Harbison *et al.* (2004). In the end, a validation set of 383 TF pairs was constructed for evaluation. We evaluated the performance of the predicted results by counting the number of true positives (TPs) among the lists of top-ranked interacting pairs and the calculated precision scores, defined by the value of TPs over the summation of TPs and false positives.

2.5 Collection of known regulatory motifs for performance evaluation

To evaluate the quality of the motifs discovered by simTFBS, we collected known motifs from literatures and database for comparison. The motifs discovered by two recent studies (Badis *et al.*, 2008; Zhu *et al.*, 2009) based on PBM experiments were included. In addition, we also collected annotated motif matrices of yeast from the MYBS database (Tsai *et al.*, 2007) (104 motifs) to enlarge the list of known motifs. In total, 307 motifs for 170 TFs were included in the final list of known motifs.

3 RESULTS AND DISCUSSION

The proposed method, simTFBS, discovered 221 TF–TF interactions as high confidence pairs (with MI score > 0.005) based on the ChIP-chip data of 203 TFs provided by Harbison *et al.* (2004). We first compared its performance with three naïve methods. After that, simTFBS was compared with two similar studies to show its advantages. A small analysis was further designed to confirm that the good performance of simTFBS was truly from the discovered motifs. Finally, we constructed a TF interacting network by incorporating the predicted interactions with the annotated interactions from literature. By examining this network, some interesting observations were addressed and discussed.

3.1 Comparison of simTFBS with three naïve methods

A fixed *P*-value cutoff is commonly used to determine the target genes of a TF when using ChIP-chip data. Here, two fixed cutoff values were used to determine the potential target genes, and the gene

sets were then used to predict TF–TF interactions by calculating the MI scores. It is noted in the caption of Figure 3 that naïve methods produced hits of higher orders of magnitude than simTFBS. The situation that the naïve methods predicted much more pairs might be owing to the fact that the number of TF pairs for prediction is large and it is likely for some of them to pass statistical test by random. On the other hand, simTFBS requires at least one similar (shared) motif to be present among the reported motifs of the two TFs. By this way, the number of potential interacting TF pairs to be considered decreased dramatically.

It can be seen in Figure 3 that using a stringent cutoff (e.g. Pc = 0.001) might exclude the information of low-affinity targets which results in limited identification of true TF–TF interactions when compared with simTFBS. On the other hand, using a loose cutoff (e.g. Pc = 0.01) might retrieve low-affinity target genes and also include a great amount of false targets, resulting in false predictions of TF–TF interactions. For the Pc-hybrid method described in Subsection 2.3, the results in Figure 3 indicated that Pc-hybrid generally performs better than Pc-0.001 and Pc-0.01 within the region where the figures were plotted. However, fewer TPs were discovered by Pc-hybrid than Pc-0.001 among the list of top-100 ranked TF pairs. On the other hand, simTFBS, which combines motif discovery with the concept of Pc-hybrid, successfully improves the predicted results when compared with the three naïve methods.

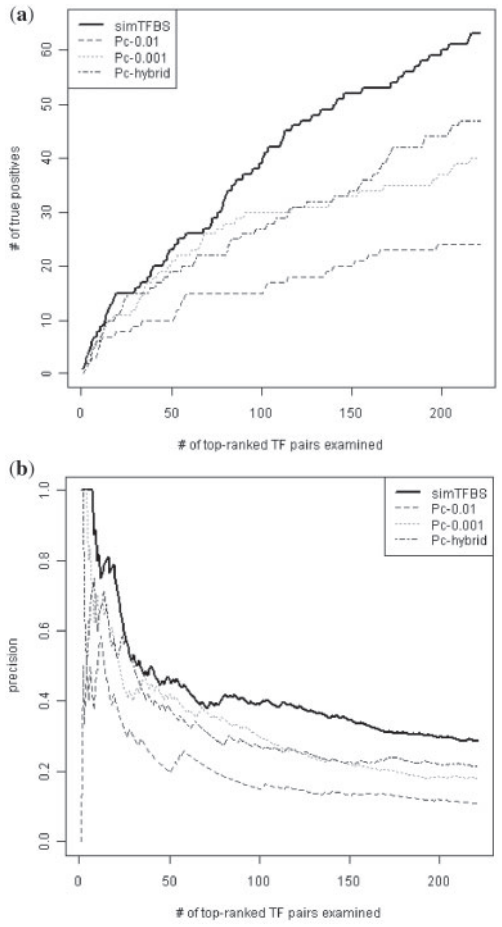

Fig. 3. Comparison of the proposed method with naïve methods of determining target genes for a TF. The validation list containing 383 interacting pairs were used to evaluate the predicted 221 associated TF pairs when compared with three naïve methods. (**a**) Accumulated TPs in the ranking list of each method; (**b**) precision evaluation. [The total numbers of high confident TF pairs (MI > 0.005) a method finally predicts are: simTFBS: 221; Pc-0.01: 1499; Pc-0.001: 980; Pc-hybrid: 1797.]

To see whether the performance of simTFBS is significantly different from the other three naïve methods, we converted the number of TPs into proportions ranging from 0 to 1 and performed a Kolmogorov–Smirnov (KS) test. The results from KS test corroborate with our hypothesis that simTFBS is superior to the other methods with extreme significance (Supplementary Table S1). In addition, we performed one-sided Fisher's exact test to see whether using different methods (simTFBS versus the compared one) cause significant changes in the distributions of true and false positives among the 221 predicted pairs (Supplementary Table S2). The highest significance was observed on the difference between simTFBS and Pc-0.01, followed by Pc-0.001 and then Pc-hybrid. In Supplementary Table S2, we further grouped the numbers of true and false negatives in gradient of the 50 top-ranked TF pairs and re-performed the Fisher's exact test in different groups. It is then observed that simTFBS has superior performance over Pc-0.01 among the 1∼150 top-ranked TF pairs, while the superiority of simTFBS over both Pc-hybrid and Pc-0.001 was observed among the 51∼100 and the 101∼150 top-ranked TF pairs, respectively. These results demonstrated that simTFBS is able to assign true interacting TF pairs with better scores by refining the target gene sets through the motif discovery process. In addition to the naïve methods considered here, a GSEA-based (Gene set enrichment analysis; Subramanian *et al.*, 2005) method was also considered as a potential method without exploiting motif information for comparison. The GSEA-based method has similar performance with the naïve approaches adopted here. The relevant materials and discussions can be found in the Supplementary Material.

3.2 Comparison with two recent studies

The first study for comparison is the work presented by Datta *et al.* (Datta and Zhao, 2008), where the authors proposed a method to dynamically determine the threshold for collecting the target gene set of a TF based on the EM algorithm. In the study of Datta *et al.*, the derived gene lists were then analyzed by log-linear models to infer cooperative binding. The predictions based on log-linear models were evaluated in the same way as was done for simTFBS in Figure 3, denoted as 'EM-LLM' in Figure 4. In addition, we also conducted the comparison of combining the dynamically determined thresholds by the EM approach with the MI score [Equation (1)] adopted in this study to predict TF–TF interactions (denoted as 'EM-MI' in Fig. 4). It is shown in Figure 4 that simTFBS outperforms both EM-LLM and EM-MI with significant improvements. The related statistical tests are provided in Supplementary Table S3 and S4.

The second study for comparison is the work presented by Yu *et al.* (2006), Motif-PIE. Motif-PIE also utilizes a motif discovery procedure when predicting interacting TF pairs. It collected known target genes from literature or databases and employed Lee et al.'s ChIP-chip data (Lee *et al.*, 2002) to infer common target genes and then predict TF–TF interactions. Though it was claimed in Motif-PIE that 152 TFs were employed to make predictions, we only

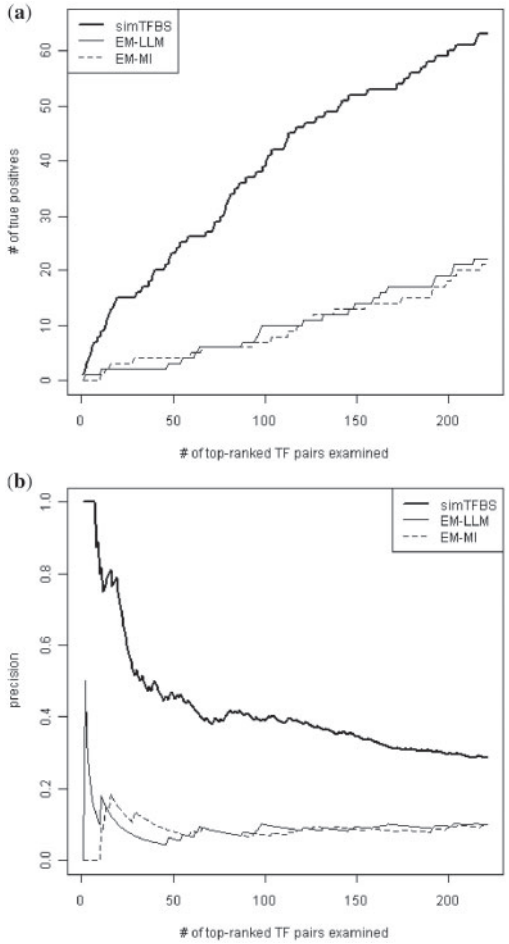

Fig. 4. Comparison of simTFBS with the method provided by Datta *et al.* Comparison was based on the 203 TFs from Harbison *et al.* (**a**) Accumulated TPs in the ranking list of each method; (**b**) precision evaluation. (Total number of TF pairs a method finally predicts: simTFBS: 221; EM-LLM: 7072; EM-MI: 4172.)

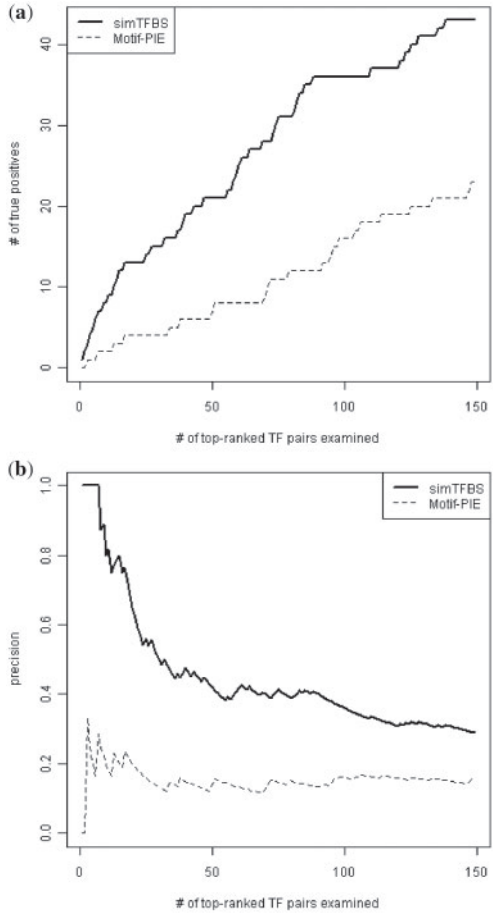

Fig. 5. Comparison of simTFBS with Motif-PIE. Comparison was based on the set of common TFs that both methods used. (**a**) Accumulated TPs in the ranking list of each method; (**b**) precision evaluation. (Total number of TF pairs a method predicts: simTFBS: 149; Motif-PIE: 282.)

successfully acquired a list of 126 TFs based on the descriptions of data preparation provided in Motif-PIE. Among the 126 TFs, 115 TFs are in common with the 203 TFs used in this study. In this regard, the comparison is based on the 115 TFs. The comparison shown in Figure 5 reveals that simTFBS significantly outperforms Motif-PIE in predicting TF–TF interacting pairs. The related statistical tests are provided in Supplementary Table S3 and S4.

3.3 Effect of recruiting *de novo* motifs

To investigate whether the novel motifs discovered by simTFBS indeed improve the accuracy of TF–TF interaction prediction, we re-conducted the predictions based on only the currently available TF binding motifs with literature support, instead of invoking a motif discovery tool. The 307 motifs were compared in a pair-wise manner and high similarity (>0.8) was reported and considered as the prediction for TF–TF interactions. This method discovered 24 TPs, considerably fewer than the 63 TPs delivered by simTFBS. The result reveals that the discovered TFBSs successfully help to identify more known interacting TF pairs and also improve the precision of the methodology.

3.4 Support from literature or microarray data

Among the 221 TF pairs with high confidence, 63 TF–TF interactions are supported by literature or database annotation. The predicted interactions were also examined from another aspect, to see if they can be explained by including one or more TFs. We say an interaction A–B can be reasonably explained if there exists another TF C, where the interactions A–C and B–C are supported by literature. There are 42 predicted interactions falling in this category. Furthermore, we examined if a predicted interactions A–B can be linked with two other TFs C and D, such that A–C, C–D and D–B are supported by literature. It turns out that there are 37 interactions fitting this criterion. In total, 142 predicted interactions can be directly or indirectly explained by literature.

We further carried out Pearson's product moment correlation coefficient test using the 'cor.test' function in R project (Development Core Team of R, 2005) to investigate the degree

of co-expression for each of the predicted TF pairs. The test was performed on the TF pairs in the predicted list and in the validation lists, respectively, using the time course microarray data performed during cell cycle (alpha factor arrested) from Spellman *et al.* (1998). The correlation coefficients and the corresponding *P*-values for the predicted 221 pairs were provided in Supplementary Table S5. In Supplementary Fig. S1, we compared the proportion of the co-expressed pairs over the total pairs in different groups (the predicted, the annotated and random) under several cutoffs of *P*-values from correlation coefficient test. The result shows that the predicted group behaves similarly to the annotated group, and both are superior to the average of 100 permutations of randomly selected TF pairs.

In addition to finding literature or expression data support for the predicted TF pairs, we also evaluated the biological meanings of the discovered motifs by comparing them with the annotated TF binding motifs, since the proposed method also outputted motifs along with each TF interaction. These motifs play an essential role in the predicting procedures of simTFBS as they were used to refine the target gene sets. In this regard, the motifs along with each TF interaction were compared with annotated TF binding motifs. It is observed that >60% (146 motifs) of the discovered motifs are similar (similarity score >0.8) to known motifs. The information of annotated motifs could be used to partition the predicted pairs into four groups (Supplementary Fig. S2). There are 22 interactions (Group 1) that are predicted based on a motif similar to both of the TFs' annotated motifs (21 cases) and two motifs similar to each of the TF's annotated motif, respectively (one case, STE12 and TEC1). For 62 interactions (Group 2), the associated motif is similar to the annotated motif of only one of the TF pair. On the other hand, there are another 62 interactions (Group 3) predicted based on a motif similar to the annotated motif of another TF rather than the predicted TF pair. Finally, we have 75 interactions categorized as 'Group 4', of which the associated motifs are novel.

We subsequently examined the precision of each group based on the validation list of TF–TF interactions, and it is observed that the precision of different groups changes dramatically. The highest precision (77.27%) was observed in 'Group 1', considerably superior to the other three groups ('Group 2': 38.71%, 'Group 3': 19.35% and 'Group 4': 9.33%). It is noticed that the precision was dramatically declined to <10% when the information of annotated regulatory motifs was insufficient. This coincidence suggested that the TF–TF interactions categorized in 'Group 1' might be easier to be discovered (and thus are enriched in the collected validation list) than TF–TF interactions associated with novel motifs (Group 4). Two further observations also support this suspicion. First, the proportion of TF pairs with co-expression profiles is observed to be higher in 'Group 1' than in the other three groups (Supplementary Table S6). Second, we observed that the motifs that are similar to known TF binding motifs are more enriched in the high-affinity targets of corresponding TFs in 'Group 1' when compared with the other three groups (Supplementary Table S7). Both observations reveal that the TF pairs categorized in 'Group 1' are easier to be discovered and thus have been largely studied. Therefore, the additional information provided by the other three groups might greatly help to enlarge the knowledge for TF–TF interactions. In addition, many TFs without annotated TF binding motifs are found to have the motifs of other TFs in their ChIP-chip data, resulting in some TPs in 'Group 2'. Moreover, by *de novo* motif discovery, simTFBS could discover a TF–TF interaction mediated by another TF out of the predicted interacting pair, resulting in some TPs in 'Group 3'.

3.5 Investigation of the role of the predicted interactions in influencing the network structure of annotated interaction network

We conducted network analysis to investigate the role of the predicted interactions in influencing the network structure of the existing interaction network (383 annotated interactions). The clustering coefficient for estimating modularity (Watts and Strogatz, 1998) was employed to analyze the new network. The definition of clustering coefficient was described in the Supplementary Material.

This analysis was executed by in-house perl programs utilizing 'Graph' module. We investigated whether the clustering coefficient of the network is improved by incorporating the predicted 221 pairs with the 383 annotated interactions (the new network contains 541 pairs in total, which include the 383 annotated interactions and additional 158 predicted pairs). A permutation test was applied and the details are provided in the Supplementary Material. It is shown in Supplementary Fig. S3 that the clustering coefficient was significantly increased when the predicted TF–TF interactions were added (from green line to red line), whereas randomly generated interactions in general decrease this measurement. This result suggests that the TF–TF interactions help to improve the modularity of the currently annotated interacting network, which might be still largely incomplete.

3.6 Limitations of simTFBS

Since simTFBS heavily relies on the performance of the motif finding step, in some cases the procedure might be unable to discover TF-pairs due to the failed predictions of motif discovery. It was observed in some TFs that the known binding motifs were not found by eTFBS. For example, the binding motif of MIG1 was not found in the list of the 10 motifs reported for the ChIP-chip data of MIG1. Though, the GAL4 binding motif was found in the ChIP-chip data of GAL4, it was not enriched in the high-affinity targets of MIG1. This resulted in the failure of predicting the interaction between GAL4 and MIG1.

In addition, it was observed that many predictions were made based on a single shared motif (i.e. only one motif of the TFs in the pair was enriched in both TFs), instead of more than one motif as expected (i.e. both motifs of the TFs in the pair were enriched in both TFs). There are some potential reasons to explain why many of the predictions were made based on a single shared motif. First, the predicted TF pairs might be in the relationship of piggy back, e.g. INO4 and OPI1. Second, it is possible that one of the TFs does not have an annotated binding motif yet. So, we did not know whether it was found or not. Third, it is the failure of motif finding algorithm that did not discover the expected motif for one of the TFs. In fact, it is not the objective of simTFBS to find the complete set of binding motifs associated with the interacting TF pairs. It is shown in this study that the information of shared motifs improves the accuracy of predicting TF interactions, and one such motif should be enough to achieve this goal.

There might be some potential biases associated with the design of simTFBS. We have provided relevant discussion regarding this issue in the Supplementary Material. In addition, though the proposed method theoretically can predict all the interacting types elaborated

in Figure 1, it cannot tell the type of mechanisms without additional information such as DNA-binding domains on the TFs. In this regard, it is suggested to include more information regarding direct or indirect binding derived by comparing ChIP and PBM data (Gordân *et al.*, 2009) or information of enrichment of biological context to refine the target gene sets in order to further improve the accuracy of predicting TF interactions in future studies.

4 CONCLUSION

In this article, we proposed a new method that incorporates *de novo* motif discovery in the procedures of analyzing ChIP-chip data. The discovered motifs were employed to refine the lists of target genes, which further improve the accuracy of predicting interactions between TFs by estimating the degree of target overlap between two TFs more correctly. The evaluation conducted in this study reveals that the proposed method, simTFBS, outperforms three naïve methods and two recent studies. The predicted TF–TF interactions are shown to improve the modularity of the currently annotated interacting network which might be still largely incomplete.

ACKNOWLEDGEMENTS

The authors also wish to thank Jen-Hao Cheng and Krishna B.S. Swamy for their valuable suggestions.

Funding: National Science Council of Republic of China, Taiwan contracts (98-2221-E-002-137-MY2 and 99-2627-B-002-004). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Conflict of Interest: none declared.

REFERENCES

Badis,G. *et al.* (2008) A library of yeast transcription factor motifs reveals a widespread function for RSC3 in targeting nucleosome exclusion at promoters. *Mol. Cell*, **32**, 878–887.

Balaji,S. *et al.* (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.

Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.

Berger,M.F. *et al.* (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat. Biotechnol.*, **24**, 1429–1435.

Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.

Bulyk,M.L. (2007) Protein binding microarrays for the characterization of DNA-protein interactions. *Adv. Biochem. Eng. Biotechnol.*, **104**, 65–85.

Bulyk,M.L. *et al.* (1999) Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nat. Biotechnol.*, **17**, 573–577.

Bulyk,M.L. *et al.* (2001) Exploring the DNA-binding specificities of zinc fingers with DNA microarrays. *Proc. Natl. Acad. Sci. USA*, **98**, 7158–7163.

Chang,Y.H. *et al.* (2006) Identification of transcription factor cooperativity via stochastic system model. *Bioinformatics*, **22**, 2276–2282.

Chen,C.Y. *et al.* (2008) Discovering gapped binding sites of yeast transcription factors. *Proc. Natl. Acad. Sci. USA*, **105**, 2527–2532.

Datta,D. and Zhao,H. (2008) Statistical methods to infer cooperative binding among transcription factors in *Saccharomyces cerevisiae*. *Bioinformatics*, **24**, 545–552.

Development Core Team of R. (2005) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Eden,E. *et al.* (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput Biol.*, **3**, e39.

Gordân,R. *et al.* (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res.*, **19**, 2090–2100.

Harbison,C.T. *et al.* (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Kar,S. and Adhya,S. (2001) Recruitment of HU by piggyback: a special role of GalR in repressosome assembly. *Genes Dev.*, **15**, 2273–2281.

Kato,M. *et al.* (2004) Identifying combinatorial regulation of transcription factors and binding motifs. *Genome Biol.*, **5**, R56.

Lee,T.I. *et al.* (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Lemmens,K. *et al.* (2006) Inferring transcriptional modules from ChIP-chip, motif and microarray data. *Genome Biol.*, **7**, R37.

MacIsaac,K.D. *et al.* (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.

Manke,T. *et al.* (2003) Correlating protein-DNA and protein-protein interaction networks. *J. Mol. Biol.*, **333**, 75–85.

McCord,R.P. *et al.* (2007) Inferring condition-specific transcription factor function from DNA binding and gene expression data. *Mol. Syst. Biol.*, **3**, 100.

Mukherjee,S. *et al.* (2004) Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.*, **36**, 1331–1339.

Narlikar,L. *et al.* (2006) Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, **22**, e384–e392.

Pilpel,Y. *et al.* (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.

Ren,B. *et al.* (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309

Linhart,C. *et al.* (2008) Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. *Genome Res.*, **18**, 1180–1189.

Spellman,P.T. *et al.* (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273–3297.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Tsai,H.K. *et al.* (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl. Acad. Sci. USA*, **102**, 13532–13537.

Tsai,H.K. *et al.* (2007) MYBS: a comprehensive web server for mining transcription factor binding sites in yeast. *Nucleic Acids Res.*, **35**, W221–W226.

Wang,Y. *et al.* (2009) Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data. *Nucleic Acids Res.*, **37**, 5943–5958.

Watts,D.J. and Strogatz,S.H. (1998) Collective dynamics of 'small-world' networks. *Nature*, **393**, 440–442.

Wu,W.S. *et al.* (2007) Identifying regulatory targets of cell cycle transcription factors using gene expression and ChIP-chip data. *BMC Bioinformatics*, **8**, 188.

Yu,X. *et al.* (2006) Genome-wide prediction and characterization of interactions between transcription factors in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **34**, 917–927.

Zhu,C. *et al.* (2009) High-resolution DNA-binding specificity analysis of yeast transcription factors. *Genome Res.*, **19**, 556–566.