BIOINFORMATICS

Vol. 19 no. 15 2003, pages 1917–1926
DOI: 10.1093/bioinformatics/btg347

Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data

Jiang Qian[1], Jimmy Lin[1], Nicholas M. Luscombe[2], Haiyuan Yu[2] and Mark Gerstein[2,]*

[1]Department of Ophthalmology, Johns Hopkins Medical School, Baltimore, MD 21287, USA and [2]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

Received on May 1, 2003; revised and accepted on July 15, 2003

ABSTRACT

Motivation: Defining regulatory networks, linking transcription factors (TFs) to their targets, is a central problem in post-genomic biology. One might imagine one could readily determine these networks through inspection of gene expression data. However, the relationship between the expression timecourse of a transcription factor and its target is not obvious (e.g. simple correlation over the timecourse), and current analysis methods, such as hierarchical clustering, have not been very successful in deciphering them.

Results: Here we introduce an approach based on support vector machines (SVMs) to predict the targets of a transcription factor by identifying subtle relationships between their expression profiles. In particular, we used SVMs to predict the regulatory targets for 36 transcription factors in the *Saccharomyces cerevisiae* genome based on the microarray expression data from many different physiological conditions. We trained and tested our SVM on a data set constructed to include a significant number of both positive and negative examples, directly addressing data imbalance issues. This was non-trivial given that most of the known experimental information is only for positives. Overall, we found that 63% of our TF–target relationships were confirmed through cross-validation. We further assessed the performance of our regulatory network identifications by comparing them with the results from two recent genome-wide ChIP-chip experiments. Overall, we find the agreement between our results and these experiments is comparable to the agreement (albeit low) between the two experiments. We find that this network has a delocalized structure with respect to chromosomal positioning, with a given transcription factor having targets spread fairly uniformly across the genome.

Availability: The overall network of the relationships is available on the web at http://bioinfo.mbb.yale.edu/expression/echipchip

Contact: Mark.Gerstein@yale.edu

*To whom correspondence should be addressed.

INTRODUCTION

Understanding of transcriptional regulatory networks is crucial in the understanding of fundamental cellular processes, such as growth control, cell-cycle progression, and development, as well as differentiated cellular function such as hormone secretion and cell–cell communication (Alberts *et al*., 1994). On a fundamental level, transcription determines when and which genes are expressed. The determination of factors that control expression can offer further insight into the misregulated expression that is common in many human diseases (Tupler *et al*., 1999; Ly *et al*., 2000).

Much research has been done related to transcription factors (TFs): some have tried to identify TFs in genomes using different methods, such as through sequence similarity or structural comparisons (Riechmann *et al*., 2000a,b; Wingender *et al*., 2001). Given known TFs, others have tried to find their binding motifs in the regions upstream of genes (Roulet *et al*., 1998; Krivan and Wasserman, 2001; Grabe, 2002; Halfon *et al*., 2002). For a TF whose binding motif is known, some researchers have started to predict gene targets of transcription factors using genome-wide sequence searches of promoter regions (Schuldiner *et al*., 1998; Zhu *et al*., 2002). Lastly, others have tried to determine targets of a transcription factor whose binding motif is unknown (Kel *et al*., 2001; Tan *et al*., 2001). This final area is the research we pursue here.

The determination of target genes of TFs has been done with different approaches. The most popular method is probably ChIP-chip, which combines the techniques of chromatin immunoprecipitation and microarray hybridization. DNA that binds specifically to a TF is purified and amplified. Genomic target loci are identified by comparative hybridization of the immunoprecipitated and control DNA probes to a DNA microarray. In yeast researchers have used this method to identify the targets of TFs such as Gal4, Ste12, MBP and SBP (Ren *et al*., 2000; Iyer *et al*., 2001).

In this work, we want to identify the targets of TFs using computational approaches. We focus on mining

gene-expression data since these data provide a direct measurement of the transcriptional program in the cell. Past analyses of microarray data have focused on clustering genes with similar expression profiles to predict protein function and interaction (Eisen *et al*., 1998; Gerstein and Jansen, 2000). However, the gene expression relationship between a TF and its targets is complex. In most cases, they do not have a correlated expression profile over a timecourse (see below). Sometimes, in fact, there is a lag time between the expression of the TF and its target (Qian *et al*., 2001).

To tackle this problem we employed support vector machines (SVMs). SVMs are a form of supervised machine learning. They use a training set to learn in advance which gene pairs have a regulatory relationship (Vapnik, 1998). The first gene in a pair is a TF, while the second is the target gene it potentially regulates. After the training stage, the machines determine probabilities for each TF–target pairing and these probabilities, with appropriate thresholds, can then be used to construct parts of a regulatory network.

This work is focused on the budding yeast *Saccharomyces cerevisiae*. Recent work has estimated that yeast has 6128 genes and 209 transcription factors (Riechmann *et al*., 2000a,b; Snyder and Gerstein, 2003). Given this, we have potentially 1 280 752 (i.e. 209×6128) combinations. Our task is to find which pairs among these 1 280 752 represent a true regulatory relationship.

METHODS

Support vector machines

In order to determine the relationship between TFs and their targets, we use SVMs. In general, the SVM is a standard supervised machine-learning algorithm, based on recent developments in statistical learning theory (Vapnik, 1998). It is designed for pattern recognition and regression and used in fields such as writing recognition, text categorization, and image classification (Vladimir and Vapnik, 1995; Joachims, 1998).

The SVM builds a hyperplane separating positive examples and negative examples in multiple-dimensional space. Unfortunately, most real-world problems involve non-separable data for which there does not exist a hyperplane that successfully separates the positive from the negative examples. One solution to the inseparability problem is to map the data into a higher-dimensional space and define a separating hyperplane there. This higher-dimensional space is called the feature space. A kernel function of the dot product of the vectors is used to avoid representing the space explicitly. For details of SVM, please refer Burges, 1998; Vapnik, 1998.

The SVM creates the separating hyperplane from the labeled training data that can then be used for prediction. Given that there are a large number of TFs with known targets to form a training set, the SVM represents an appropriate algorithm for regulatory network prediction.

Here we use an implementation of SVM by Brown *et al*. (2000). Our focus is not in developing the SVM methodology but seeing the degree to which it can be applied to gene expression data.

Encoding of gene expression data

To encode our regulatory network prediction problem in a form suitable for training SVMs, we construct TF–target pairs. These pair a *known* transcription factor R and a putative target gene T that may be regulated by this factor. For instance, the pairing $(R \Rightarrow T)$ means transcription factor R regulates gene T. To connect this pairing with expression information, we note that each gene in the pair is characterized by a set of expression experiments, which comprise data from samples collected at various time points during the diauxic shift, the mitotic cell cycle, sporulation, and heat shock (Spellman *et al*., 1998; Gasch *et al*., 2000). In total, we used 79 gene expression data points to characterize each gene. Then putative TF–target pairing corresponds to a 158-element gene expression vector, in which the first 79 expression data points are for the (TF) while the second 79 are for the regulated gene.

Positive training examples

Positive examples were obtained from two transcription databases: TRANSFAC (Wingender *et al*., 2001) and SCPD (Zhu and Zhang, 1999). These two databases bring together information from the biochemical literature on TFs and their regulated genes. In this study, we only include sequence-specific TFs and exclude general TFs, such as the RNA polymerases and the TATA-binding protein. In total, we used 175 TF–target pairings as positive examples.

Negative training examples

As with other supervised machine learning methods, negative examples are needed to train properly. In our case, a negative example would be a gene pair that we know definitely has no regulatory relationship. Note that this is distinct from a gene pair about which we have no positive information. Unfortunately, there are essentially no papers on definitive negative relationships in the biochemical literature. Consequently, we employed a number of strategies to come up with appropriate negative examples.

In the onset, one can easily make negative examples in a number of ways. For example, two genes encoding ribosomal proteins would have no regulatory relationship between (though they may, of course, be regulated by the same factor). Another possibility is creating two artificial gene-expression profiles using randomized numbers. However, while easy to construct, such examples may not be optimal for machine learning. In principle, SVMs find the boundary between the positive and negative examples. If the negative examples are made too different from the positive examples, the learned boundary is loose and thus it would be problematic to detect subtle cases.

In the end, we constructed negative examples in two ways: (i) for the TFs with known binding sites, we searched for these sites genome-wide in the upstream regions of all genes. Then for target gene T whose upstream sequence contains no binding site for transcription factor R, the pairing $R \not=> T$ constitutes a negative example. (ii) For TFs whose binding sites are unknown, we randomly select another gene to construct a negative example. To make sure that the randomly selected gene is not regulated by the TF, the expression profile of the second gene is permuted while keeping the expression profile of the TF constant.

In total, we constructed 1750 negative examples for training, which is 10 times the number of positive examples. The reason for this ratio between the positive and negative examples will be explained below.

The imbalance problem

In machine learning, when there is great disparity between the size of the positive and negative training sets, one must take into consideration a training difficulty called the imbalance problem (Japkowicz, 2000; Japkowicz and Stephen, 2002). This problem occurs when there is a large difference between positive and negative examples of the data. In such a situation, the algorithm will accurately predict the over-represented class, but its prediction of the under-represented class will mostly be incorrect. In the extreme case, the under-represented class will be ignored. For example, for a positive to negative ratio of 1 : 1000, an algorithm that always predicts negative will be correct 1000 times and incorrect only once. There are two approaches towards overcoming the imbalance problem. (i) Increasing the size of the under-represented set by random resampling and (ii) decreasing the size of the over-represented set by random removal of its members (Japkowicz, 2000; An *et al*., 2001; Japkowicz and Stephen, 2002).

The imbalance problem is encountered in our TF target prediction since (we believe) there are definitely more negative transcriptional relationships than positive ones. For the yeast genome, even if one assumes that each TF regulates ∼200 genes, there would be a 1 : 30 ratio between positive and negative examples. [These numbers are reasonable given the numbers from some of the recent ChIP-chip experiments (Horak *et al*., 2002; Lee *et al*., 2002)].

The imbalance problem also has implications for the relationships between threshold, coverage and error rate. (After fully developing our method, we illustrate some of these issues by showing the different error rates and coverage values for 1 : 1 and 1 : 10 training sets in Fig. 3.)

Restricting the prediction to the subset from yTAFNET

In order to alleviate the imbalance problem, we decreased the prediction set from all possible TF–target pairings (i.e. $1\,280\,752 = 209 \times 6128$) to just the pairings suggested by the yTAFNET database (Devaux *et al*., 2001).

We used an initial set of potential TF–target gene pairs obtained from the yTAFNET database. This database combines 72 published experiments and extracted the up-or down-regulated target genes associated with different TFs in different states. In most of the experiments in this database, the TFs were knocked out and the genes selected had significant changes in their expression. Note, these genes are not necessarily the *direct* target of the TF, but they are more likely to be the targets than randomly selected genes from the whole genome. We hoped this would reduce the imbalance between the positive and negative examples. Since this is a preliminary set, the selection criteria did not have to be stringent and thus we chose the 1.5 fold set from yTAFNET, which showed genes that were up-or down-regulated at least 1.5 fold. We selected 36 TFs for prediction. This resulted in 46 059 putative TF–target pairings that we assessed using our SVM.

RESULTS

Expression relationship between TF and targets is not simultaneous

We assessed the problem of prediction of transcription targets based on their expression profiles. Figure 1A–D shows four examples of expression profiles between TFs and their regulatory targets. The black lines are the expression profiles for TFs while the red lines are the corresponding regulated genes. At first glance, one can see there are no obvious relationships between the expression profiles of a TF and its regulated gene.

Looking closer, it seems that there exist some relationships between the expression profiles. For example, In Figure 1A, from conditions 10 to 20, they have a simultaneous relationship, while from conditions 44 to 60, the two profiles display an inverted relationship. In Figure 1B, from conditions 52 to 62, the two profiles show that the target gene has a shifted response compared with the TF.

In Figure 1D, from conditions 45 to 62, the expression profile of the target gene is an exaggerated profile of the TF. However, one cannot calculate the significance of these relationships. Especially, when these four positive examples are compared with the four negative ones (Fig. 1E–H), in which the two expression profiles do not have a regulatory relationship.

To get a global view of the problem, we calculated correlation coefficients between the expression profiles of TFs and their corresponding target genes for both the positive and negative examples in the training set. The distributions, shown in Figure 2, are quite broad, ranging from -0.2 to 1. It is clear that one cannot predict the regulatory relationship purely from the correlation of the expression profiles between the TF and its target gene. Interestingly, the distribution for the positive examples displays shoulders both to the left

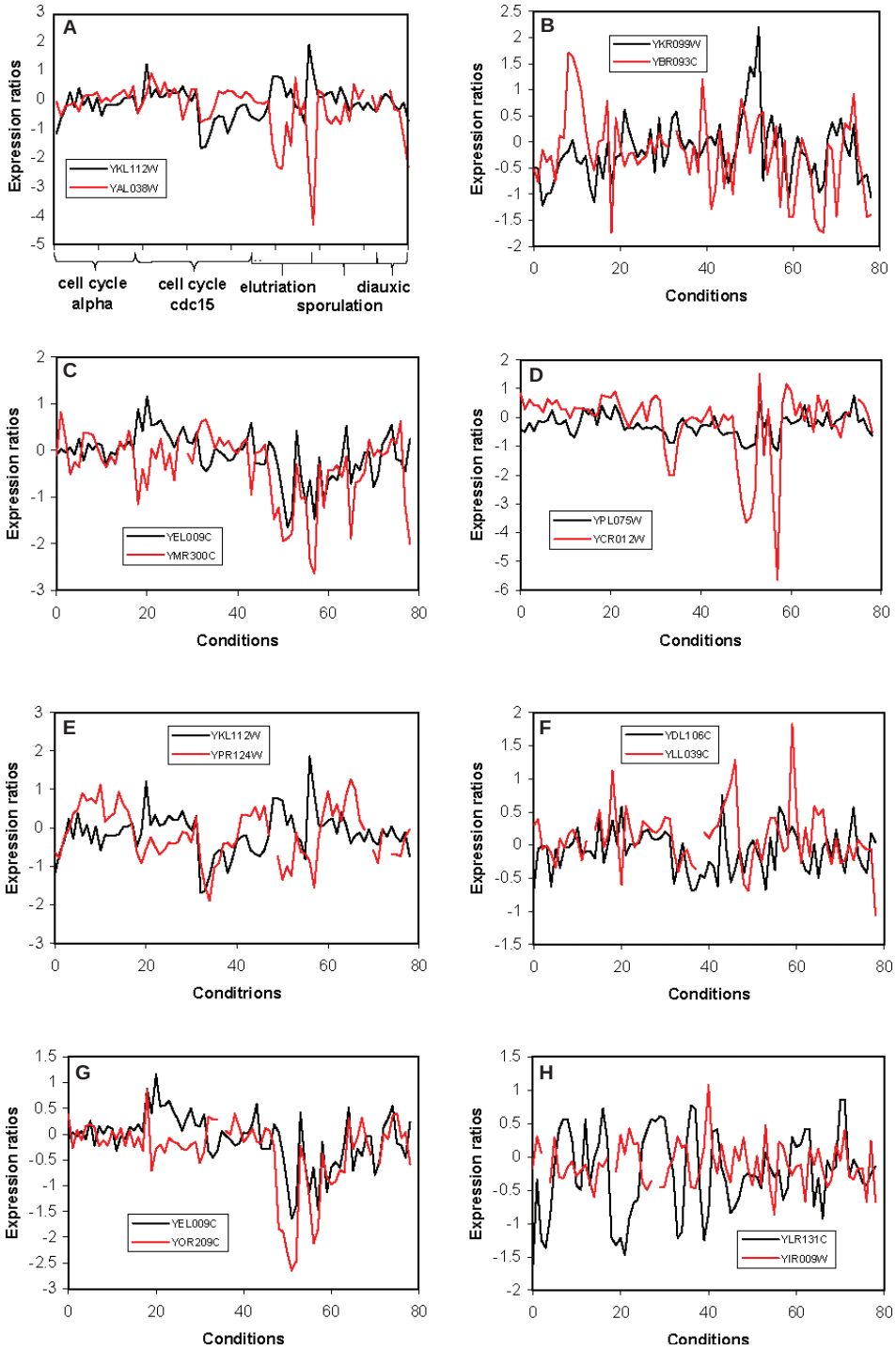

Fig. 1. Expression profiles of TF and target pairs. Sample expression profiles showing different control relationships are shown in this figure. The TF profiles are shown in black and the gene target in red. Sections (**A**)–(**D**) show known positive relationships while sections (**E**)–(**G**) show known negative relationships. (A) YKL112W controls YAL038W almost directly for the first half and inversely for the second half. (B) YKR099W controls YBR093C with a time shift relationship between points 50 and 60. (C) YEL009C seems to control YMR300C inversely from points 20 to 40 but directly from 40 to 70. (D) YPL075W seems to control the slope of YCR012W from 40 to 60. (E) YKL112W seems to have a mixed inverse and direct relationship with YPR124W throughout the profile. (F) YDL106C seems to have a general correlation with YLL039C on a macroscopic scale, but the detailed changes are very different. (G) YEL009C seems to have broad correlations with YOR209C, perhaps controlled by similar processes, but there is very low correlation of the details. (**H**) YLR131C has no clear relationship with YIR009W.

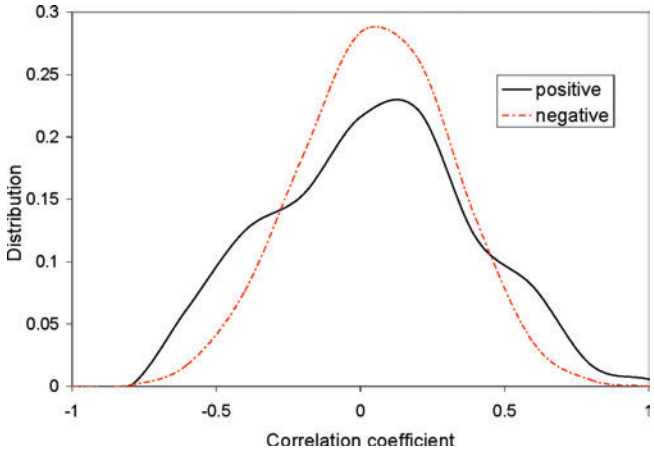

Fig. 2. Correlation-coefficient distributions. In order to determine general relationships between TFs and their targets, we calculated the distribution of correlation coefficients of the known positive examples compared with the distribution based on negative relationships. The distribution of positive correlations is shown in a solid line and shows two shoulders; the distribution of negative correlations is shown in a dotted line and has a near Guassian distribution.

and right of the main peak. This means that one has more chance to find positive relationships than negative relationship if two expression profiles show high correlation or high anti-correlation.

Evaluating the performance in cross-validated fashion

While we can see that simple correlations are not sufficient to predict the regulatory relationship, the gene expression profiles should contain the information necessary to determine regulatory networks. However, this information is rather subtle. Machine learning approaches are useful here, since they can find subtle relationships that are not immediately apparent and require no explicit description of the connection between the input information and predicted relationship.

In this work 175 positive and 1750 negative examples were used for evaluation of the performance of SVM. Each example consists of a pair of genes and is characterized by 158 gene expression levels in different experimental conditions. The performance of the SVM was evaluated by three-fold cross-validation. In other words, 117 positive and 1170 negative examples were used for training and the rest of the examples for prediction. The random split between the training and prediction sets was repeated 10 times and the average performance was calculated. Table 1 shows the results of cross-validation using five different kernel functions. The sensitivity can be calculated as $\mathrm{Sn} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, while the specificity is $\mathrm{Sp} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$. (The symbols

Table 1. Three-fold cross-validation using five different kernel functions

	TP	TN	FP	FN	Sensitivity	Specificity	Precision
Power = 1	29	467	113	29	0.50	0.81	0.78
Power = 2	36	536	44	22	0.62	0.92	0.90
Power = 3	32	561	19	26	0.55	0.97	0.93
Power = 4	22	568	12	36	0.38	0.98	0.92
Radial	9	579	1	49	0.16	1.00	0.92

For each kernel function (powers 1–4 and radial), true positives, false positives, true negatives, false negatives, sensitivity, specificity, and precision are shown in the different columns. The methods of calculation are described in the text.

TP, TN, FP and FN are defined the number of true positive, true negative, false positive and false negative obtained from the prediction, respectively.)

The accuracy describes overall performance and is defined as $A = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})$. One can see that the accuracies for powers 3 and 4 and radial kernel functions are similar. Power 3 is slightly better than others; the accuracy rate for this kernel function is 93%, and this value provides an evaluation of the overall prediction quality including positive and negative predictions.

Since the majority of the predictions are from the negative samples, a more strict evaluation of the prediction is the precision $[\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})]$, which concentrates on the sample of predicted positives. As 32 out of the 51 predicted positives are, in fact, true positives, the precision of the prediction is 63%.

The threshold for the prediction: ROC graph

We also calculate the relationship between the prediction coverage and the error rate. The prediction coverage is the percentage of the true positives in the real positives (i.e. the sensitivity Sn.) The error rate E is percentage of the false positives in the real negatives (i.e. $E = 1 - \mathrm{Sn}$). It is easy to imagine that both the prediction coverage and error rate increase with the decreasing threshold. If one wants to include as many true positives as possible, in the mean time, much more false positives will occur in the prediction. Normally one needs to find the optimal point that has the minimal amount of wrong predictions. However, in our case, we are more interested in the low error rate than in the high coverage. In other words, the quality of the prediction is more important than the coverage.

In Figure 3 the coverage versus the error rate is shown for our prediction. This graph is in the standard form of a ROC (receiver–operator characteristic) plot. Each point on this graph represents a threshold for positive and negative classification. An optimal threshold should have high prediction coverage and a low error rate. A threshold of 0.0 was used for the further work.

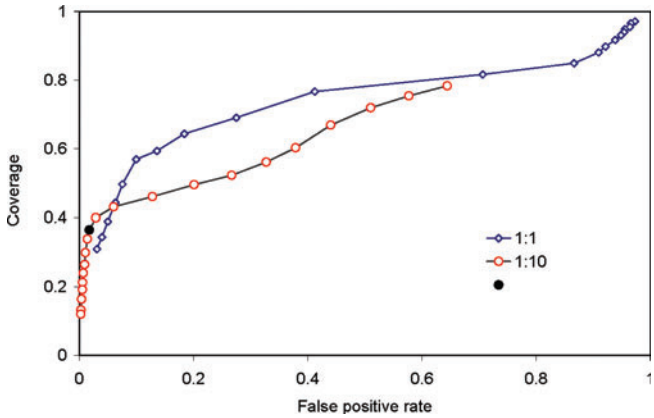

Fig. 3. ROC graph: prediction coverage versus error rate. Prediction coverage is the percentage of predicted positives that are true positives while the error rate is the percentage of predicted positives that are false positive. With a higher coverage rate, there would be an associated higher error rate. In the graph, two different plots are given, depending on the ratio of the size of the positive to negative training examples—what we call the positive-to-negative-training ratio. One plot has a ratio of 1 : 1 while the other has 1 : 10. Each point on the graph represents a different threshold setting. For the experiment, we chose a threshold setting of 0.0 with a positive-to-negative-training ratio of 1 : 10, which is shown by the darkened circle. This corresponded with a coverage rate of ∼36% and an error rate of ∼1.8%.

Genome-wide prediction of yeast transcription targets

For the genome-wide prediction of regulatory targets of yeast TFs, we used all 175 positive and 1750 negative examples as a training set. The set of 46 059 possible TF–target pairings to perform predictions on was obtained from the yTAFNET database (see methods). For 36 transcription factors, a total of 3419 TF–target pairings were found by our prediction.

Overall statistics for the predictions are presented in Tables 2 and 3. Table 2 lists these 36 TFs along with the function and number of targets they control. The average number of targets per TF is ∼93. Table 3 presents the overall statistics from another perspective. The table shows all the gene targets in the study that are controlled by 10 or more (TFs). The average number of (TFs) per target is ∼1.8.

Overall network structure

In Table 4, we show some examples of our predictions. We attempt to depict the overall network predicted in Figure 4. However, due to the large number of predicted relationships, it is only possible to show a small fraction of the total relationships in the figure. The entire network can be obtained from our website http://bioinfo.mbb.yale.edu/expression/echipchip

Table 2. TFs in the study

Transcription factor	Number of targets	Transcription factor	Number of targets
STE12	1032	SINS	37
RAP1	306	SIR2	25
ZAP1	286	SIR3	18
RTG1	271	HIR2	16
SOK2	194	GLN3	11
YAP1	189	YAP3	11
RPD3	135	MBP1	9
GCN5	105	GCN4	7
GCR1	104	SWI6	7
TUP1	71	SWI5	6
PDR1	68	ARGR1	5
PPR1	66	RGT1	4
PHO4	65	GAL4	3
SWI4	63	STB4	2
SIR4	63	YAP7	2
RPN4	59	CAT8	1
HDA1	55	TEC1	1
SSN6	47	PDR3	1

AVERAGE = 92.92

This table lists the 36 TFs used this study. For each TF, the function and the number of predicted targets are shown in the columns. The average number of targets per TF is ∼93.

Table 3. Top TF targets

Target	Number of TF	Target	Number of TF
ZRT1	20	HSP150	11
YGP1	16	ALD6	11
HXT2	15	PH05	11
PHO12	14	FAA3	10
HIS4	14	TDH3	10
FBP1	13	ASN1	10
SIP4	12	CLN2	10
ADE12	12	SUC2	10
ARG5,6	12	ILV3	10
PCK1	12	GIC2	10
HXT5	11	TYE7	10

This table shows the top TF targets that are controlled by more than 10 TFs. The average number of TFs for each target is ∼1.8.

Finally, Figure 5 shows the relative chromosomal localization of the targets of 10 TFs (randomly selected) across the genome. For the most part, there is an even distribution of targets for each factor, which corroborates with data from ChIP-chip studies (Horak *et al.*, 2002; Lee *et al.*, 2002).

Comparison with ChIP-chip results

To further evaluate our prediction, we compared our results with two recent genome-wide experiments, which determined the TF targets with the ChIP-chip approach (Horak *et al.*,

Table 4. Predicted TF–target examples

Transcription factor	TF target	Score	Transcription factor	TF target	Score
RTG1	FET3	18.83	RPD3	ALD6	6.965
RAP1	RPS1A	18.48	GCR1	FET3	6.601
SIR4	GPM1	11.27	YAP3	PGK1	6.591
SIR4	PGK1	10.66	RPN4	PDC1	6.473
RAP1	RPL40B	9.301	RTG1	HXT6	6.438
PDR1	PDC1	9.096	RPN4	PGK1	6.437
ZAP1	FET3	9.007	SOK2	ALD6	6.325
RPD3	GPM1	8.65	GCR1	ALD6	6.29
ZAP1	PGK1	8.377	RTG1	YGP1	6.222
ZAP1	GPM1	8.231	RAP1	APL3	6.011
GCR1	PDC1	8.2	RTG1	ADE5,7	6.006
ZAP1	PDC1	8.159	RAP1	RPS4A	5.994
RAP1	RPL26B	8.13	TUP1	PGK1	5.882
STE12	GPM1	8.089	RAP1	PHO12	5.867
YAP1	FET3	8.084	PHO4	RPL25	5.804
YAP1	GPM1	7.994	HIR2	TDH3	5.752
PDR1	ALD6	7.751	ZAP1	ALD6	5.747
RTG1	HXT7	7.707	TUP1	PDC1	5.711
SIR4	TDH3	7.56	RPN4	GPM1	5.685
STE12	PDC1	7.477	RAP1	RPS9A	5.625
RTG1	ACS2	7.368	PDR1	YEF3	5.556
RAP1	RPL7A	7.274	TUP1	FET3	5.484
ZAP1	ENO2	7.105	ZAP1	TDH3	5.479
SSN6	FET3	7.07	ZAP1	ACS2	5.472
GCR1	GPM1	7.02	SIR4	RPL21B	5.458

The first column is the TF, second column is its target, third column is the prediction scores. (The entire list can be obtained from our website.)

2002; Lee *et al*., 2002). In Figure 6A, we present the overlap of the TFs shared between two experimental data sets and our prediction set in terms of a Venn diagram. Note that the Horak and Lee data sets only have two TFs in common. The overlap between our prediction set and the Lee data set is 18 TFs, and there is only one common TF for both experimental data sets and our prediction set.

Based on the (relatively few) shared TFs, we analyzed the targets and TF–target relationships that were common between the experimental data sets and our predictions (Fig. 6B). In general, there is not a large overlap. Between the two experimental ChIP-chip data sets, there were only 17 common TF–target relationships, accounting for approximately 3% of all the determined relationships (where the number of determined relationships is based on the smaller data set). On the other hand, our computational predictions have an overlap of 70 TF–target relationships with Lee data set and 7 with Horak data set, which accounts for approximately 6 and 4% coverage of these data sets. There were no TF–target relationships that were consistently found in all three data sources. In summary, we found that the agreement between our results and two experiments is comparable to the agreement (albeit low) between the two experiments.

DISCUSSION AND CONCLUSION

In our analysis, we develop a machine learning approach to decipher the complex relationship between a TF and its target. Genome-scale analyses of TF targets are difficult and both experimental and computational techniques are in the processes of refinement. From our predictions, for the 36 TFs, we predict a total of 3419 targets. On average, each TF controls approximately 93 targets and each target is controlled by 1.8 TFs. This suggests that the lack of a clear relationship between TF and their targets as shown in Figure 1 can perhaps be due to the fact that most targets are not controlled by one single TF. However, the fact that one TF controls so many targets points to the importance of studying these relationships.

Other *in silico* approaches with regulatory target predictions use binding site information. However, shared tertiary structure is often the determinant for binding. This is not predicted using sequence information. Furthermore, for many TFs, binding motifs are yet to be determined. Therefore, our method provides an additional perspective that does not require as much derived information.

As with many bioinformatic analyses, there is restriction based on the initial data set, on which predictions are based. Our accuracy rate would definitely improve with incorporation of more microarray data as with the addition of more pairs of TF and targets. Furthermore, the 63% cross-validation rate with known relationships provides a measure for our analysis. However, it is important to note that this number assumes that the known relationships are accurate and do not include undiscovered, unannotated true positives. From our initial predictions, we expect coverage of 36% with an error rate of less than 2%.

The generated predictions from our analysis are useful for researchers as a preliminary target list for their TF of interest. Actual relationships need to be verified with experimental work. However, this work provides a new method of TF target prediction that will be useful with the growing amount of microarray data and knowledge of TFs. Quick predictions can be made from existing microarrary experiments and will be a useful tool as a first step in TF target prediction.

Recent studies by Lee *et al*. (2002), Ren *et al*. (2000) and Iyer *et al*. (2001) have examined the relationship between TF and their targets using the ChIP-chip approach. Our analysis examined the consequences of gene control using expression levels. However, there are only small overlaps between the different experimental data sets and with our predictions. This is most likely due to the temporal nature of TFs. For example, different TFs can compete for the same target gene. Furthermore, at different times in the cell cycle, there are differing environments with different TFs present.

As future work is done, the combination of *in vitro* and *in silico* techniques will be valuable in determining the relationship between TFs and their targets. Consensus data from different experiments will increase the fidelity of the

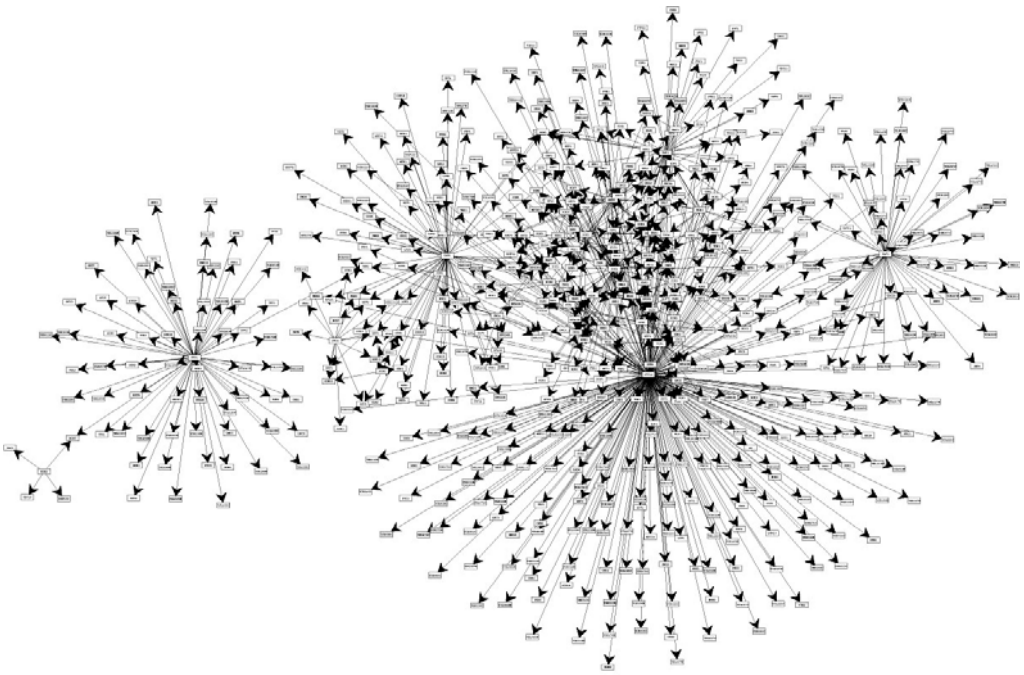

Fig. 4. Overall network. The complex interconnected network of the TFs and all their targets. Because the network is dominated by TFs targets that do not provide further control with relatively few TFs, there appears to be several centers of control with many targets.

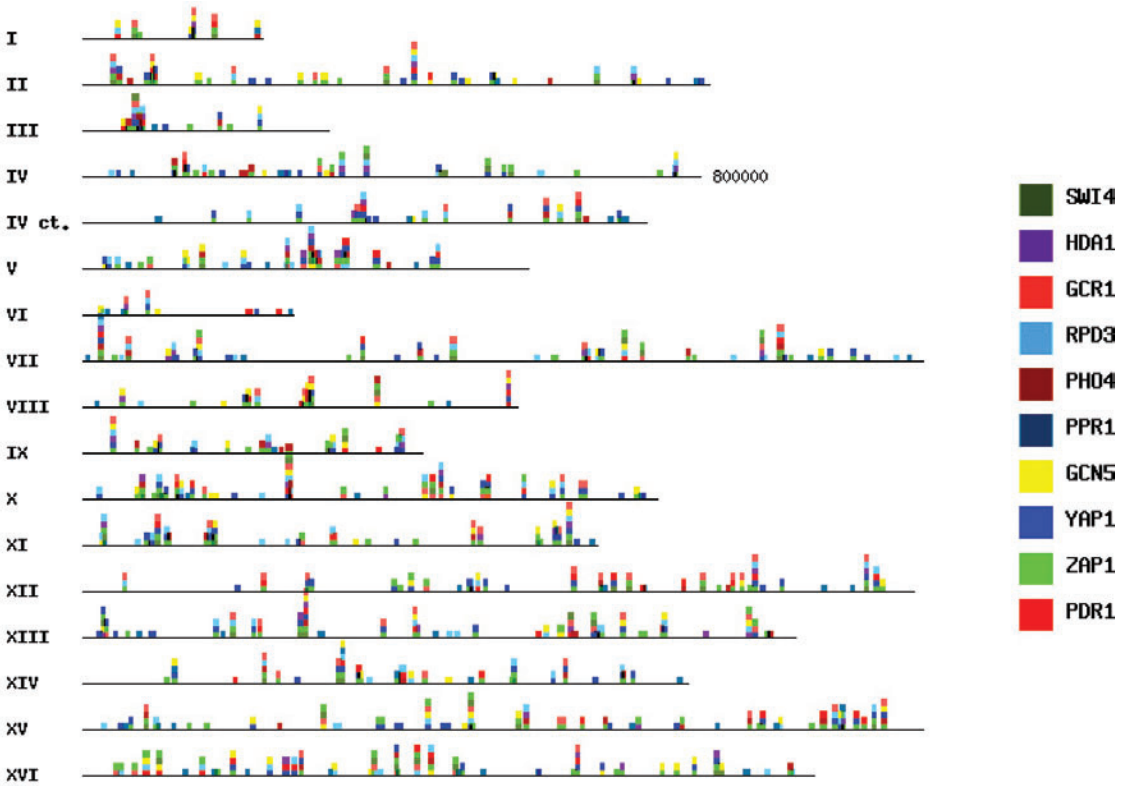

Fig. 5. Chromosomal position. Positions of genes controlled by ten TFs. For each TF, their targets are colored on the chromosome map of the yeast genome. Chromosome IV is divided into two lines: the first line contains position from 1 to 800 kb and the second shows position from 800 kb on. This provides an overall chromosome view of transcription control.

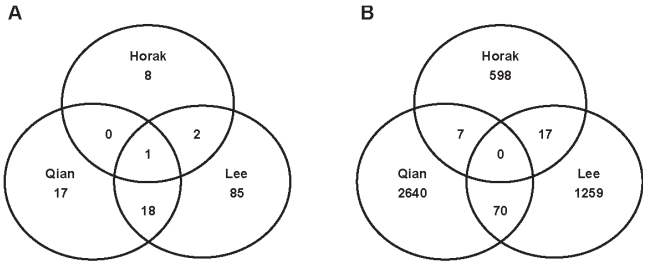

Fig. 6. Comparison of two ChIP-chip data sets with our predictions. (**A**) The sharing of the TFs that were used in the three studies. (**B**) The number of TF–target pairs that were shared among the three data sets. This only included the predictions from the data sets that shared common TFs shown in (A).

predictions. As different groups study more common TFs and with consideration of the point in cell cycle and the state of the cell, researchers will be able to better understand the control of genes within the cell. With the growing library of expression analyses and other data sources, computational techniques will provide a more complex description of the relationship.

REFERENCES

Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K. and Watson,J. (1994) *Molecular Biology of the Cell*. Garland Publishing, New York.

An,A., Cercone,N. and Huang,X. (2001) A case study for learning from imbalanced data sets. *Advances in Artificial Intelligence: Proceedings of the 14th Conference of the Canadian Society for Computational Studies of Intelligence*, **1**, 1–15.

Brown,M.P., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M.,Jr. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Burges,C.J.C. (1998) A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, **2**, 121–167.

Devaux,F., Marc,P. and Jacq,C. (2001) Transcriptomes, transcription activators and microarrays. *FEBS Lett.*, **498**, 140–144.

Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.

Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D. and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell.*, **11**, 4241–4257.

Gerstein,M. and Jansen,R. (2000) The current excitement in bioinformatics-analysis of whole-genome expression data: how does it relate to protein structure and function? *Curr. Opin. Struct. Biol.*, **10**, 574–584.

Grabe,N. (2002) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol.*, **2**, S1–S15.

Halfon,M.S., Grad,Y., Church,G.M. and Michelson,A.M. (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. *Genome Res.*, **12**, 1019–1028.

Horak,C.E., Luscombe,N.M., Qian,J., Piccirrillo,S., Gerstein,M. and Snyder,M. (2002) Complex transcriptional circuitry at the G1/S transition in *Saccharomyces cerevisiae*. *Genes Dev.*, **16**, 3017–3033.

Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.

Japkowicz,N. (2000) The class imbalance problem: significance and strategies. *Proceedings of the 2000 International Conference on Artificial Intelligence*, **1**.

Japkowicz,N. and Stephen,S. (2002) The class imbalance problem: a systematic study. *Intelligent Data Analysis*, **6**, 429–450.

Joachims,T. (1998) Text categorization with support vector machines: learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*. Springer, Berlin.

Kel,A.E., Kel-Margoulis,O.V., Farnham,P.J., Bartley,S.M., Wingender,E. and Zhang,M.Q. (2001) Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors. *J. Mol. Biol.*, **309**, 99–120.

Krivan,W. and Wasserman,W.W. (2001) A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.*, **11**, 1559–1566.

Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

Ly,D.H., Lockhart,D.J., Lerner,R.A. and Schultz,P.G. (2000) Mitotic misregulation and human aging. *Science*, **287**, 2486–2492.

Qian,J., Dolled-Filhart,M., Lin,J., Yu,H. and Gerstein,M. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions. *J. Mol. Biol.*, **314**, 1053–1066.

Ren,B., Robert,F., Wyrick,J.J., Aparicio,O., Jennings,E.G., Simon,I., Zeitlinger,J., Schreiber,J., Hannett,N., Kanin,E. *et al*. (2000) Genome-wide location and function of DNA binding proteins. *Science*, **290**, 2306–2309.

Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al*. (2000a) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R. *et al*. (2000b) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.

Roulet,E., Fisch,I., Junier,T., Bucher,P. and Mermod,N. (1998) Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. *In Silico Biol.*, **1**, 21–28.

Schuldiner,O., Yanover,C. and Benvenisty,N. (1998) Computer analysis of the entire budding yeast genome for putative targets of the GCN4 transcription factor. *Curr. Genet.*, **33**, 16–20.

Snyder,M. and Gerstein,M. (2003) Genomics. Defining genes in the genomics era. *Science*, **300**, 258–260.

Spellman,P.T., Sherlock,G., Zhang,M.Q., Iyer,V.R., Anders,K., Eisen,M.B., Brown,P.O., Botstein,D. and Futcher,B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces Cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **9**, 3273–3297.

Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res.*, **11**, 566–584.

Tupler,R., Perini,G., Pellegrino,M.A. and Green,M.R. (1999) Profound misregulation of muscle-specific gene expression in facioscapulohumeral muscular dystrophy. *Proc. Natl Acad. Sci. USA*, **96**, 12650–12654.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Vladimir,N. and Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, Berlin.

Wingender,E., Chen,X., Fricke,E., Geffers,R., Hehl,R., Liebich,I., Krull,M., Matys,V., Michael,H., Ohnhauser,R. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.

Zhu,Z., Pilpel,Y. and Church,G.M. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.*, **318**, 71–81.