Building Transcription Factor Classifiers and Discovering Relevant Biological Features

Dustin T. Holloway[1], Mark Kon[2], Charles DeLisi[3§]

[1]Molecular Biology Cell Biology and Biochemistry Department, Boston University, 5 Cummington Street, Boston, USA
[2]Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, USA
[3]Bioinformatics and Systems Biology, Boston University, 44 Cummington Street, Boston, USA

[§]Corresponding author

Email addresses:
DTH: dth128@bu.edu
MK: mkon@bu.edu
CD: delisi@bu.edu

Abstract

Background

The network of interactions between transcription factors and the genes they regulate governs many of the behaviours and responses of cells. An increasingly important goal in post-genomic research is discovering links in this network. For many transcription factors high quality sets of target genes have been identified. Given such example sets, machine learning technologies create rules for target identification. Furthermore, they facilitate the identification of specific attributes, such as particular DNA patterns or gene expression experiments which help distinguish true targets from non-targets. We have previously reported the development and validation of a new supervised learning approach to regulator site identification, and used it to predict targets of 104 transcription factors in yeast. We now include a new sequence conservation measure, expand our predictions to include 59 new TFs, and implement a ranking method to reveal the most important biological features contributing to TF binding. All together we integrate 8 genomic datasets covering a broad range of measurements including sequence conservation, sequence overrepresentation, gene expression, and DNA structural properties.

Results

Overall, the method reported here is able to predict binding sites with an accuracy of 76%, and the top 25 classifiers yield 86% accuracy. The new predictions match well with known biology of many TFs, and all predictions are available for download on a web server as well as for visualization in the VisAnt network analysis suite [1, 2]. Analysis of the regulatory network in VisAnt allows simple selection of biologically interesting network motifs such as a new feed-forward loop involving Swi6, Rfx1, and the target gene YMR279C. Other network-wide properties, such as the TFs and genes serving as hubs are examined. The most highly regulated genes are enriched for functions in carbon and energy metabolism, while the most pervasive regulators control the cell cycle, growth, and the stress response.

Using the cell cycle regulatory gene Swi6 as a case study, our results show the efficacy of robust feature ranking techniques for selecting biological attributes which are important for regulatory control. Using classifiers based on simple sequence motifs and gene expression measurements, the feature ranking for Swi6 correctly identifies the expression experiment on Mbp1 deletion mutants as being important for target regulation by Swi6. This makes sense since Mbp1 forms a complex with Swi6 during the cell cycle. Feature ranking also identifies a DNA sequence matching the known Swi6 binding site, and indicates that it is conserved and over-represented in target genes. This sequence can be seen using the SVM techniques in new Swi6 targets, including Isc1, a gene important for the biosynthesis of ceramide, a bioactive lipid currently being investigated for its anti-cancer properties. These predictions suggest possible new roles for Swi6 in lipid/ceramide metabolism.

Conclusions

SVM-based classifiers provide a comprehensive platform for analysis of regulatory networks. Post-processing of classifier results can provide high quality predictions, and robust feature ranking strategies can deliver valuable biological insight into the regulatory functions of specific transcription factors. Future work on

this method will focus on expanding the analysis to the human genome and applying similar strategies to the analysis of cancer gene expression datasets

Background

Many factors influence the regulation of genes and their protein products within the cell. Chromatin condensation, DNA methylation, and histone acetylation/methylation can affect the accessibility of a gene's cis-regulatory sites to trans-acting factors. On the RNA level, mRNA splicing, mRNA editing, microRNA silencing, and RNA degradation can all affect the ability or efficiency of translating mRNA into active protein. Nevertheless, the primary mode of regulatory control is the association of transcription factors with their binding sites in DNA. These binding sites occur most often in promoter regions, the stretch of DNA upstream of the transcription start site. The string of nucleotides bound by a particular TF is not identical at every recognized site. Instead, the TF distinguishes a flexible motif, or shared pattern of bases.

Founding work in discovering and representing binding sites involved the use of position specific scoring matrices (PSSMs) [3-6], which represent the frequency of nucleotide bases at each position in a known motif. New predictions are sites which match the PSSM based on a score threshold [3]. Methods in motif discovery seek to estimate the PSSM model, given a set of sequence regions known or hypothesized to be bound by a particular TF. Many techniques for discovering and predicting binding sites have been reported [7-14], and an evaluation of the state of the art in current motif-discovery methods is available [15].

Despite their broad usefulness, detection by PSSM is beset by a high rate of false positive predictions. Some TF matrices can produce predictions at a frequency of 1 in every 500bp [16]. Often, there is not enough information to construct high quality matrices. To improve target prediction, more sophisticated machine learning approaches can be used. Supervised learning schemes begin with more information and seek to generate classification rules based on a user-provided set of positive and negative examples. Some work has been published on supervised classification schemes for predicting TF binding targets, and we have briefly reviewed a few of these in our previous work ([17] and [18]), which focused on developing and applying a support vector machine variant to predict transcription factor binding sites in Saccharomyces cerevisiae. More specifically we compared the effectiveness of various datasets for predicting the binding of 104 TFs to their target genes, and we evaluated predicted targets based on the integration of all datasets. We now expand that work to include 163 TFs, revise our machine learning strategy to be more robust, and construct and analyze the gene regulatory network in *S. cerevisiae*. All predictions are now available online, including the full transcriptional network, which can be analyzed in the VisAnt browser [1, 2].

Genomic datasets have high dimensionality, with many numerical features often in the thousands or tens of thousands describing each gene in the dataset. Many classification algorithms, e.g., *k*-nearest-neighbors or neural networks, will perform poorly in such a setting unless selection criteria are used to drastically reduce the number of features. Support vector machines perform well with high dimensional data and have been shown to provide excellent classification accuracy with many genomic datasets (see Methods).

Positive examples for regulation are taken from known TF binding sites published in the literature. Negatives are a randomly chosen subset of those genes

found not to be bound by a TF in ChIP-chip experiments (typically these are the genes with highest *p*-values and thus least significant binding). A schematic representation of the classification workflow is presented in Figure 1. See the Methods section for a full description of classifier construction and validation. Using several genomic datasets (described below) an SVM classifier is constructed for each TF on a chosen set of features and then evaluated using a leave-one-out cross validation approach. Some difficulties remain for choosing the negative set at this stage. The selection of negatives does not guarantee that the chosen genes are truly not targets of the TF. In the worst cases where some TFs have few known targets, the classifier would contain only a few negative examples. This can introduce fluctuations in the results, biasing the performance in unpredictable ways.

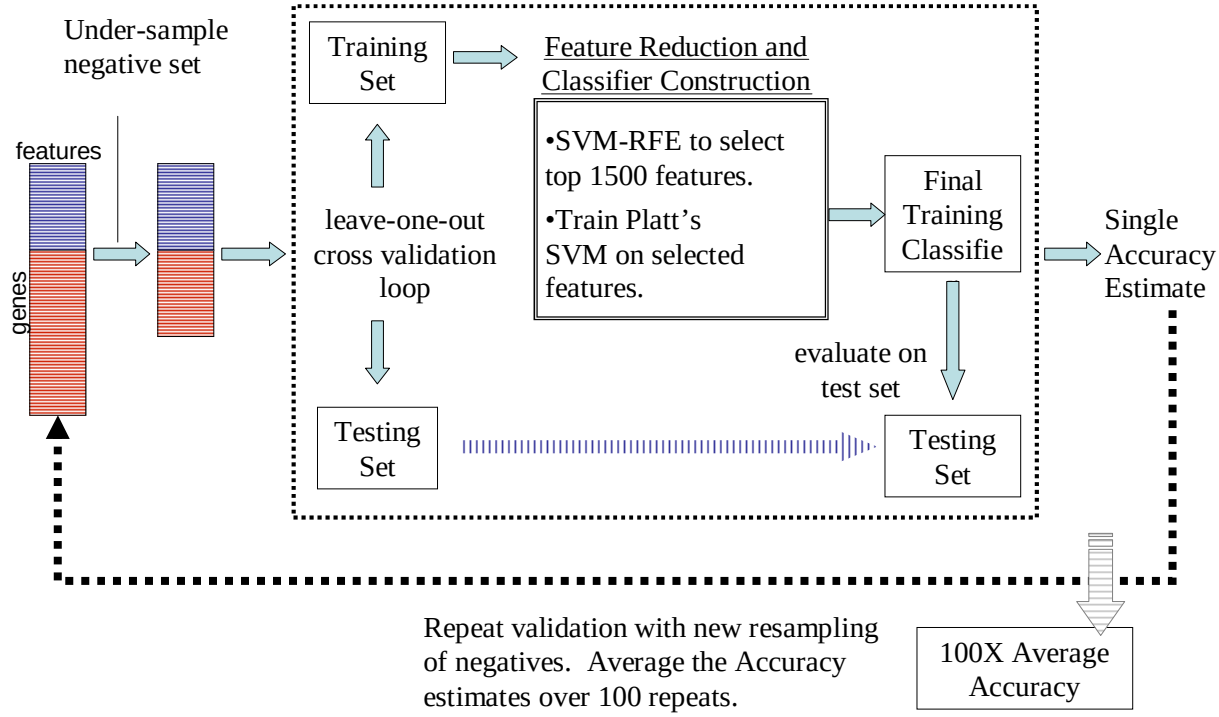


Figure 1 - SVM Framework

This figure shows the data mining scheme for making TF classifiers. 100 classifiers are constructed for each TF, each using a different random sub-sample of the negative set. First, on the far left, the negative pool for a TF is under-sampled, so that it is the same size as the positive set. A classifier built on the training set is evaluated using leave-one-out cross validation (center, dashed box). For every cross-validation split, the top 1500 features are selected using SVM-RFE and the classifier is trained and finally used to classify the test set (left out sample). This process is repeated 100 times, and the accuracy for the procedure is the average of the 100 cross-validation accuracies. To classify a potential new target for a TF, all 100 classifiers are applied to the gene's feature vector, and the average posterior probability is calculated (probability of being a target). A probability greater than 0.5 indicates a positive classification.

To resolve these difficulties we first assure that the pool of possible negatives is a minimum of 600 genes, or at least three times the size of the positive set. From this pool we randomly select a set of negatives which is equal in size to the known positives. After constructing and validating a classifier on these examples, the process is repeated one hundred times, each time with a new random resampling from the

negative pool. The average cross-validation accuracy from all repeats is reported. This assures that small-sample classifiers are challenged with a diverse set of negatives and that fluctuations due to errors in the negative set tend to average out.

Once the choices of positive and negative sets are made, genes are described by a set of attributes or features to be used by the algorithm to learn the classification rule. A gene may be described by its expression measurements over a set of conditions, or perhaps by counts of various motifs present in its promoter. For example, to capture the sequence composition of a gene's promoter, all possible nucleotide patterns of length 4 may be counted in the 800 base pairs 5' of the transcription start site. This results in a vector 256 elements long, each entry being the count of a particular 4-mer in the promoter. In our previous study, we examined 26 different types of features including various types of *k*-mer counts, expression data, and phylogenetic profiles. Here we choose 8 feature types selected to represent a diverse set of data including sequence, structure, expression, and conservation.

Several things set our current work apart. First, balanced training sets (equal numbers of positive and negatives) and the random resampling of negatives provides for more robust classifiers. Second, rather than using all features to make a classifier we apply recursive feature elimination to select those that are most relevant. Most importantly, a ranking is also created which can be used to identify the specific features that are most useful for separating target genes from non-targets. The simultaneous ranking of all features allows us to easily discover important biological aspects of regulation. Third, several of our methods have been improved. We introduce a new dataset of *k*-mer counts weighted by their conservation in alignments with sequences from closely related species. Finally, we expand the transcriptional network to include 59 new TFs (163 total) and analyze the global network properties of the regulatory interactions in yeast. This analysis highlights the transcriptional network hubs; the factors which control the most genes and the genes which are bound by the largest set of regulators. Again, Swi6 is taken as an example and a new potential feed-forward loop is discussed which may take part in the cell cycle and DNA damage response.

Results and Discussion

Parameter tuning, feature selection, and training set size

Two parameters must be set in our framework before building a classifier. One parameter, denoted by *C*, determines the amount of misclassification that is tolerated. The second is the number of features to select during training. Since it would be computationally prohibitive to choose these parameters during the training of every classifier, they are first optimized on the classifier for one TF. The learned values are then applied to the remaining classifiers.

The transcription factor YIR018W is chosen for parameter selection since it is known to regulate ~70 genes, which is close to the average for all TFs being analyzed. Features for all genomic datasets are concatenated to produce large attribute sets for each gene. SVM-RFE is used to rank each feature during classifier training (see Methods) and various feature subset sizes are tested using a leave-one-out cross validation. Thus we allow the datasets to adjust, automatically selecting the most important features, irrespective of the data sets from which they originated. Figure 2 shows the effect that changes in feature number have on classifier accuracy. Although as few as five features achieve 70% accuracy, the addition of more features

continues to improve accuracy until 1500 features are selected, where accuracy is approximately 85%. 1500 is then the number chosen for the remaining TFs.

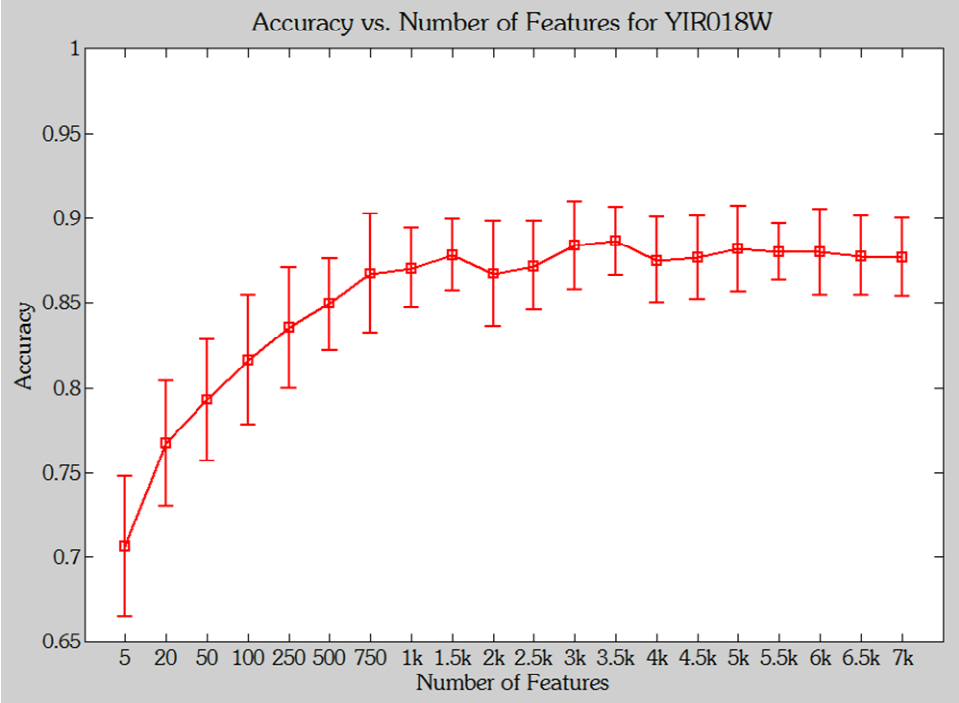


Figure 2 - Feature Elimination

YIR018W was used as a prototype to determine the number of features needed to build a classifier. This graph shows the accuracy of classifiers for YIR018W built using different numbers of features selected using SVM-RFE. Error-bars show one standard deviation of 50 classifiers constructed at each feature number using different sub-sampling of the negative set. 1500 features were selected as the point where accuracy reaches a plateau.

Using a grid search as described in Methods, we have chosen $C = 0.0078$, although results are relatively insensitive to changes in value for $C < 1$. Classifiers are then constructed for the remaining transcription factors. As discussed in Methods, performance is measured using leave-one-out cross-validation. Since 100 classifiers are trained for each TF using 100 randomly chosen negative sets, the reported accuracy is the average for 100 trials. The accuracy for all yeast binding sites (over all classifiers) is 76%; for the best 25 classifiers it is 86%. All predictions are in the form of a class conditional probability as described in [19]. New predictions are the result of averaging the assigned probabilities from each of the 100 classifiers for a TF. An average probability greater than 0.5 indicates a positive classification. Higher confidence predictions can be obtained by increasing the threshold. Throughout this manuscript probabilities given with the capital *P* refer to the mean posterior probability assigned by the SVM classifiers (see Methods). Lower-case *p* refers to *p*-values assigned using other statistical tests.

Classifier accuracy is loosely correlated with the size of the positive set, where TFs with more known targets tend to have more accurate classifiers (Supplementary Figure 1). This implies that classifier performance could improve in the future, as more experimental targets are discovered. This effect is seen mainly for classifiers

with few positives. Indeed, many classifiers with 20 or fewer positive examples perform poorly.

Web Server, Network Visualization and Analysis

One of the origins of network structure is combinatorial regulation; i.e. single TFs often regulate more than one target, and particular targets are often regulated by more than one TF. Such networks can be augmented with data on protein-protein interactions and displayed as a repertoire of connections (the cell's network repertoire), different subsets of which are selected by particular environments. Here we display the underlying repertoire using the VisAnt analysis and visualization system [1, 2]. The repertoire is available at the VisAnt website and can be accessed in the methods table as "TFSVM."

Visualization in VisAnt allows comparison of predictions to be integrated with many other large scale genomic datasets including protein-protein interaction, gene expression, GO functional annotation, and genetic interaction. VisAnt also includes a sliding bar which allows adjustment of the network based on a threshold for accepting a predicted association. Thus predictions can be embedded in networks mined at any specified degree of stringency for further analysis.

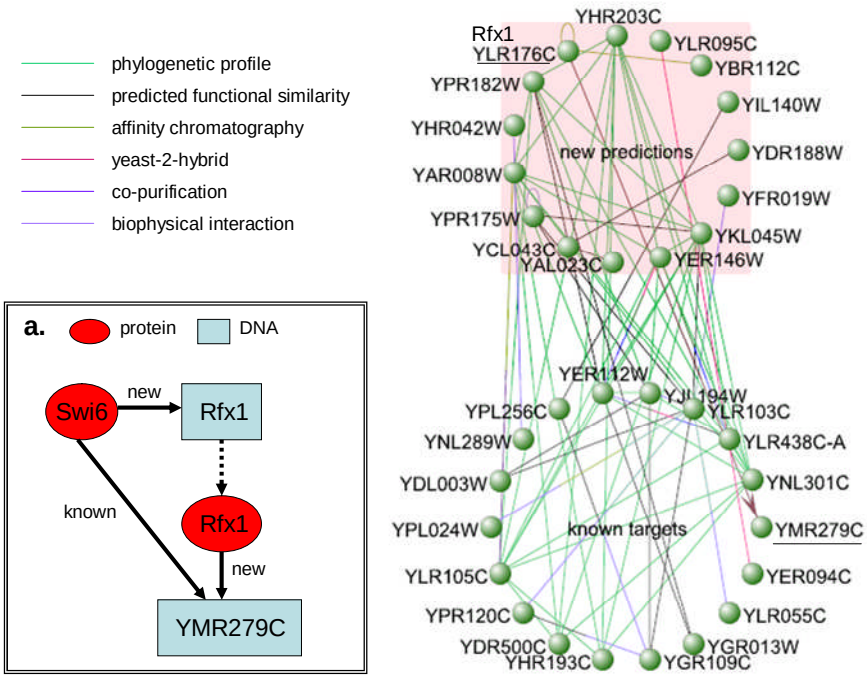


Figure 3 - Partial Regulatory Network of Swi6

A sub-network showing some of the new predictions for Swi6 and how they are interconnected to previously known targets, using the VisAnt browser. Rfx1 and YMR279C are underlined in the network. 3a shows and up-close schematic of the feed-forward loop between Swi6, Rfx1, and YMR279C.

Figure 3 displays a sub-network of known and newly predicted ($P > 0.95$) targets of Swi6. Only a portion of the known Swi6 targets are shown. This subset was chosen

because the genes are as interlinked with each other as they are with newly predicted targets. The new predictions are highly connected to known targets by a variety of experimental and computational methods, including phylogenetic profiling, genetic interaction, yeast two-hybrid, Bayesian predicted functional similarity, co-purification, affinity chromatography, and synthetic-lethal experiments. Both gene groups in Figure 3 contain cell cycle genes, and the most common connection is phylogenetic profiling, suggesting considerable functional similarity. The network perspective supports the prediction of common regulation of these highly interacting genes, and makes it easier to formulate testable hypothesis about the relationships of regulatory targets.

It is clear from Figure 3 that a new target of Swi6, YLR176C(Rfx1), is a transcription factor which regulates a previously known Swi6 target, YMR279C. This arrangement is a feed forward loop, suggesting that YMR279C is under strict combinatorial control by these two factors (Figure 3a). It should be noted that although our method independently predicts that Rfx1 is a regulator of YMR279C. Rfx1 was reported to bind the promoter of YMR279C in an early ChIP-chip study[20]. An updated analysis of those results by the same group removed YMR279C from the dataset (can be downloaded from [21]), and a subsequent ChIP-chip experiment did not show significant regulation[22]. Due to this ambiguity the Rfx1-YMR279C interaction is not in our positive set; nonetheless, it is predicted by the classifier for Rfx1.

Rfx1 is a repressor known to be involved in the cell-cycle DNA damage checkpoint. Inactivation of Rfx1 in response to cellular DNA damage or replication block causes the induction (i.e., de-repression) of many genes[23]. As further evidence that Rfx1 is indeed regulated by Swi6, expression of the two transcription factors was examined during the alpha-factor arrested cell cycle time course[24]. Figure 4 shows the expression of Swi6, Rfx1, and two reference genes which show expression peaks in G1 and S-phase. Across the 18 experiments in the time course, the two factors show a correlation coefficient of 0.6.

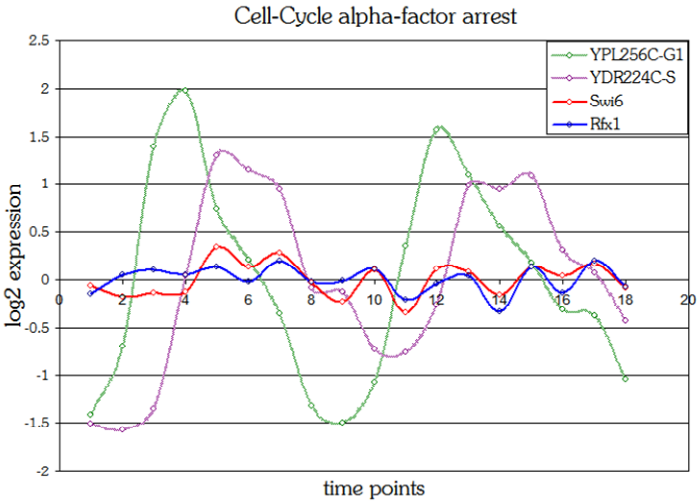


Figure 4 - Swi6 and Rfx1 in the Cell Cycle

This graph shows the Log2 expression values of Swi6 and a newly predicted target, Rfx1 (also a TF) during the cell cycle. G1 and S-Phase are marked by the expression of two prototype genes, YPL256C and YDR224C, which are known to have peak expression in G1 and S-phase, respectively. Swi6 and Rfx1 show correlated expression during the cell cycle.

Since Swi6 is known to be important for the G1/S transition, the expression in these specific experiments was examined more closely. Using prototypical G1 and S phase reference genes, eight time points spanning the peak of G1 through the peak of S phase were selected. In these eight experiments Swi6 and Rfx1 show a stronger correlation of 0.73, which is statistically significant at $p = 0.02$. Interestingly, Swi6 binds several genes in the DNA damage response pathway including Dun1(known), Rad53(new, $P = 0.96$), and Mec1(new, $P = 0.76$). These targets are upstream kinases known to phosphorylate Rfx1 in the DNA damage response pathway[23].

Finally, it has been shown that Rad53 directly phosphorylates Swi6, delaying progression of the cell cycle into S-phase[25]. The ultimate target of the new feed foreword loop is YMR279C, which is an uncharacterized gene showing sequence similarity to membrane transport proteins. This gene shows a 9-fold induction in response to DNA damage[26]. Taken together, this evidence suggests that Swi6 is crucial to the DNA damage response at the end of G1, and that YMR279C plays a role in DNA damage response and perhaps cell cycle progression. Deletion mutants of YMR279C are viable[27] but it is not known how this deletion affects the cell cycle or DNA damage response. Examining these mutants for deficiencies in growth and DNA damage response may shed some light on the true function of YMR279C. Detailed experiments including reporter assays would be needed to determine how closely interlinked Swi6, Rfx1, and YMR279C are on the transcriptional level. As a working hypothesis, it appears that Swi6 is available to activate DNA damage response genes such as Rfx1, ensuring they are present at crucial times in the cell cycle. Normally Rfx1 is repressing its targets and YMR279C will not be activated. In times of DNA damage Rfx1 is inactivated by a phosphorylation cascade allowing YMR279C and other response genes to be induced by Swi6, resulting in cell cycle arrest.

In any regulatory network it is of interest to know which genes are most heavily under transcriptional control, and which TFs exert the most control by regulating large numbers of genes. When analyzing global network properties, we limit our analysis to high quality predictions by only including TFs that have a classification accuracy greater than 0.6 (there are 130 such TFs), and targets that have a true positive probability ≥ 0.95. In the resulting network, many genes are under strong regulatory control. For instance, 125 genes are regulated by 12 or more transcription factors. These genes show statistical enrichment ($p \leq 0.05$) in several GO biological process categories. The enriched categories are mainly involved in carbon metabolism and energy generation (see Supplementary Table 1), which are crucial functions expected to be under intense control. Other important processes include DNA damage checkpoint, DNA recombination, DNA damage response, acetyl-CoA catabolism, NADP(H) metabolism, and telomere maintenance. In all, 13 transcription factors show very broad regulatory control. These 13 TFs regulate more than 300 genes each at high significance levels ($P \geq 0.95$). This set of TFs includes the pervasive regulators Abf1 and Reb1, as well as TFs involved in the cell cycle, growth, and stress response (see Supplementary Table 2).

The full set of predictions for 163 TFs in S. cerevisiae are available at http://cagt10.bu.edu/TFSVM/Main%20Frame%20Page.htm . Users may query a

transcription factor, returning the list of predicted targets, or a gene, returning a list of possible regulators. The class conditional probabilities described above may also be set by the user, providing an adjustable threshold on the confidence of predictions for each TF. In addition, the cross validated accuracies for all classifiers have been posted online, as well as the top 50 ranked features for each TF-classifier.

Prediction Analysis and Feature Rank

As described in Methods, the features for each classifier are ranked using SVM-RFE. Ranked features can be used to reveal interesting biological aspects of regulation and suggest directions for future experiments. We again use Swi6 for a case study. Swi6 interacts with Mbp1 and Swi4 during the cell cycle (G1/S transition) and in meiosis [28, 29]. This TF has 142 known targets and its SVM classifier has a prediction accuracy of 83%. The known targets of Swi6 and the new predictions (at true positive probability thresholds of $P > 0.5$ through $P > 0.95$) are significantly enriched ($p < 0.05$) in the expected GO biological processes including cell cycle, regulation of cell cycle, mitotic cell cycle, DNA repair, etc. Some of the new categories for which targets at $P > 0.95$ show enrichment include chromatin assembly/disassembly (p = 1e-10), septin checkpoint (p = 2.7e-3), and lipid metabolism (p=8.4e-4). These new targets fit with the current intuition about Swi6 regulation and suggest possible new roles of action.

Further regulatory implications can be seen by examining the importance of the features used for classification. Since feature ranking is performed on every training set during cross validation, and because the cross validation is repeated 100 times with random negative sets, there is a total of 14200 feature rankings available (142 examples times 100 cross validation repetitions = 14200 rankings). Using this ensemble of rankings, features are re-ranked based on the frequency with which they appear in the top 40 rank-ordered features. This allows the selection of attributes that are robust to changes in training data and can serve as reliable markers for differentiating regulatory targets. Figure 5 shows a plot of the features for Swi6 sorted by their occurrence in the top 40 ranked features within the 14200 rankings. Only a relatively small number of features retain high importance across a majority of the rankings. The first 10 features are in the top 40 of more than 65% of the rankings while the remaining features fall off sharply in reliability.

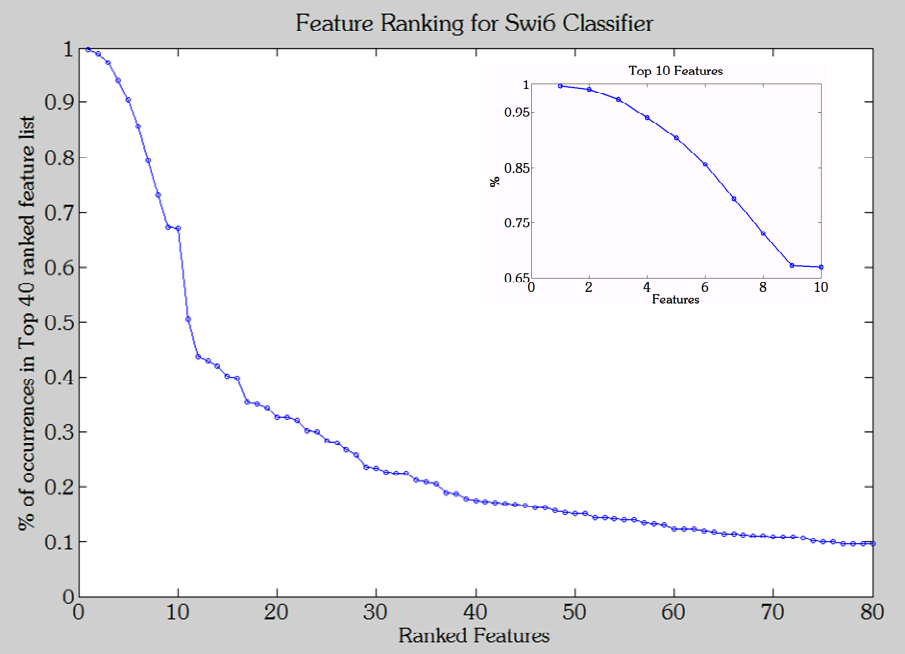


Figure 5 - Swi6 Feature Ranking

A feature ranking is created on every training set during cross validation of a classifier. Since 100 classifiers are made for every TF, this results in hundreds of separate rankings (100 * #Training Points). This plot shows the frequency with which a feature shows up in the group of top 40 features. When sorted by this frequency, only the most important features remain at the top of the list. For Swi6 the first 10 features are in the top 40 of 60% of the rankings.

The most important feature for identifying known targets of Swi6 is a microarray experiment measuring expression of genes in an Mbp1 deletion mutant. This makes sense since Swi6/Mbp1 interact and function as coactivators at many promoters. By *t*-test the observed expression change is significant between the negative set and the known positives (p = 3.7e-25), and between the negative set and the predictions made at 95% confidence (p = 9.14e-27, 280 genes). For more details on how microarray data are incorporated as features in the datasets see Methods.

The next 4 highest ranked features identify the *k*-mer ACGCG/CGCGT as being important for classification by conservation and overrepresentation. The *k*-mer overlaps highly with known binding sites for Swi6; for example, Swi6 binds CGCGAAA in the Cln2 promoter[30]. The overrepresentation of this sequence in the positive training set can be seen by examining the calculated *E*-values, which are the scores used to determine the significance of each *k*-mer (see Methods). Viewing this score in the negative set genes as compared to the positives is informative, and Figure 6 plots the distribution of *E*-values in the genes of the negative set, positive set, and the predicted targets at $P > 0.95$. The graph was generated by placing the genes in each set (i.e., positive, negative, and predicted) into 5 equally spaced bins based on *E*-value. The known and predicted targets clearly show enrichment of ACGCG as compared to the negative genes.

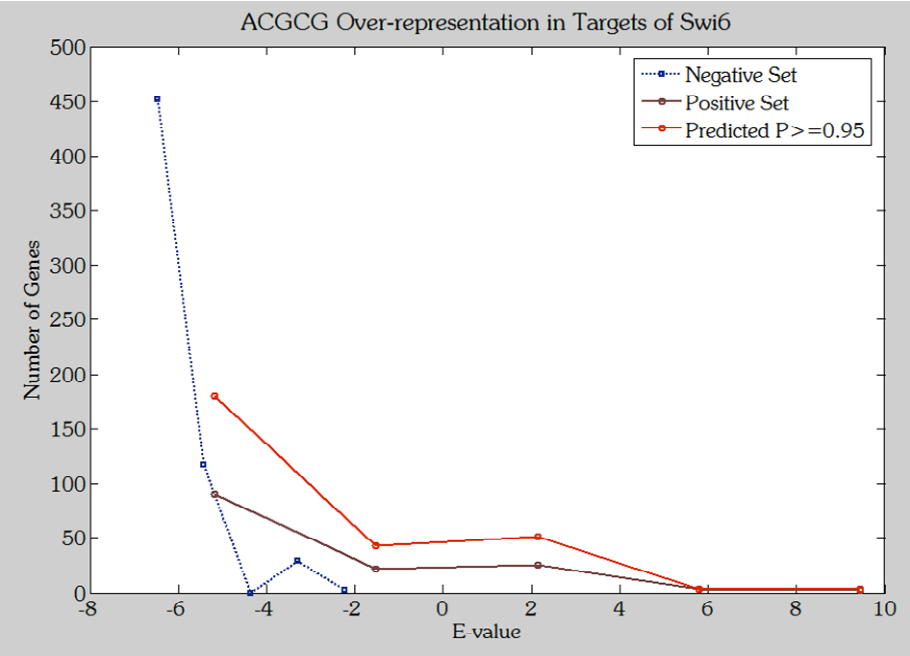


Figure 6 - Overrepresentation of ACGCG in Swi6 Target Promoters

This plots the number of genes versus the E-value of ACGCG in groups of target promoters. Three categories of promoters are shown: (i) negative set promoters in blue, (ii) positive set promoters in violet, (iii) predicted targets at $P \geq 0.95$. For each category, genes are grouped into 5 equally spaced bins based on the E-value of overrepresentation of ACGCG. The center locations of those bins are plotted on the *x* axis and the number of genes in each bin is on the *y* axis. Positive and predicted target promoters of Swi6 show higher overrepresentation of ACGCG than negative set genes.

The conservation of ACGCG is ranked more highly than overrepresentation, indicating that this sequence is preserved in promoter alignments which include 7 *Saccharomyces* species. Figure 7a shows such an alignment in the promoter region of the Isc1 gene, a newly predicted target of Swi6 (Figure 7b shows a similar alignment in the Sur2 promoter). Two occurrences (highlighted in red) of the indicated *k*-mer appear in close proximity in a highly conserved segment of the Isc1 promoter. Isc1 is an important gene since it is the only member of the extended family of sphingomyelinases (SMases) which is present in the yeast genome. This SMases are responsible for generating ceramides, which are bioactive lipids known to modulate a variety of cellular processes including cell growth, senescence, apoptosis, and the cell cycle[31]. They also contain a newly discovered P-loop-like domain, which is conserved in the SMase family from yeast to humans[32]. Isc1 is the closest yeast homolog to the human neutral SMase2 gene and has thus become the prototype for exploring the functions of this enzyme class.

a

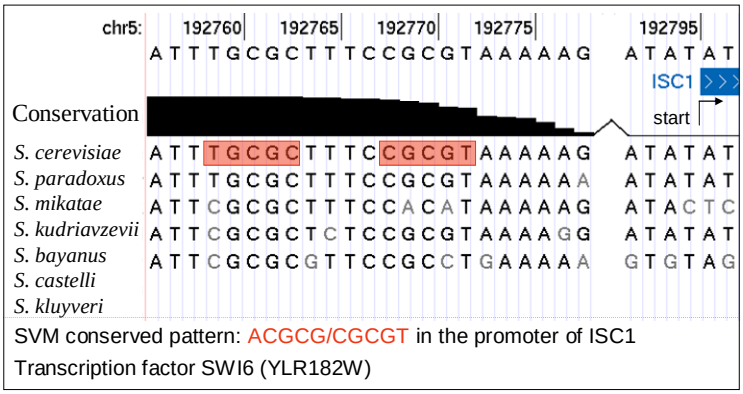


Figure 7 - Conservation of ACGCG in Swi6 Target Promoters

This figure shows multi-genome alignments of selected promoter regions of Swi6 targets taken from the USC Genome Browser. Instances of ACGCG and its reverse complement are highlighted in red. The conservation score is the output of the PhastCons algorithm. The score is a posterior probability assigned to each nucleotide, giving the likelihood that the nucleotide resides in a conserved element, as defined by a hidden Markov model of "slow" and "fast" evolution. A) alignment in the promoter region of ISC1, showing two instances of the ACGCG motif. B) alignment in the promoter region of Sur2, also showing two instances of the conserved motif.

Isc1 is a new prediction, not previously annotated as a target of Swi6. The regulation of Isc1 is now being actively explored, and it was recently shown that Isc1 is linked to cell growth in yeast[31] as well as being important for fermentative growth and sexual reproduction[33]. Perhaps more importantly, recent experiments in human and mouse models have demonstrated ceremide enhanced cancer cell death, indicating that ceramide could act synergistically with other chemotherapeutic agents[34]. Indeed, several therapeutic compounds are currently under development which modulate ceramide metabolism. The prediction that Isc1 is regulated by Swi6(Mbp1) is significant since it provides a direct link between cell cycle regulation and the generation of bioactive lipids via the hydrolytic pathway. Further investigation will be necessary to determine the biological significance of this link and whether the human ortholog, SMase2, shows a similar connection to the cell cycle. It is possible that ceramide production via SMase2 could serve as a target for anti-cancer therapy which can be easily studied in yeast models. It would be of interest to create sphingomyelinase inhibitors which would help in dissecting the specific roles of the ceramide biosynthetic pathway (SMase independent) and the hydrolytic pathway (SMase dependent) in regulating ceramide levels and cell death. The information that predicted targets of Swi6 include SMase and are enriched in genes functioning in cellular lipid metabolism (GO category *p*-value=8.4e-4) suggests that *Swi6 has a greater role than previously appreciated in controlling lipid metabolism and its coupling to cell growth, apoptosis, and reproduction.*

As a further example of the usefulness of SVMs coupled to a robust feature ranking, we briefly explore the results for the TFs Gzf3 (YJL110C), and Ash1 (YKL185W). The feature ranking for the factor Gzf3 (YJL110C) indicates that the melting temperature at positions -274 to -286 in target promoter regions is different than that in non-target promoters. Figure 8 shows a plot of the average melting temperature in sets of yeast promoters averaged within a moving 20bp window. Known targets clearly have a reduced melting temperature at the identified positions as compared to negative set or average genes. The relationship is still present in the targets which show a true-positive probability >0.95. This group contains 72 targets, 27 of which are new predictions. Although it is unclear how the melting temperature and helix stability in this region affects regulation by Gzf3, it is possible that Gzf3 or other factors induce changes in DNA compaction or stability which alter regulation at these promoters. Binding sites of Gzf3 do not appear to be concentrated in this region, implicating the activity of other factors at this site. In any case, feature ranking has identified specific nucleotides which can be tested experimentally for their role in transcriptional regulation.

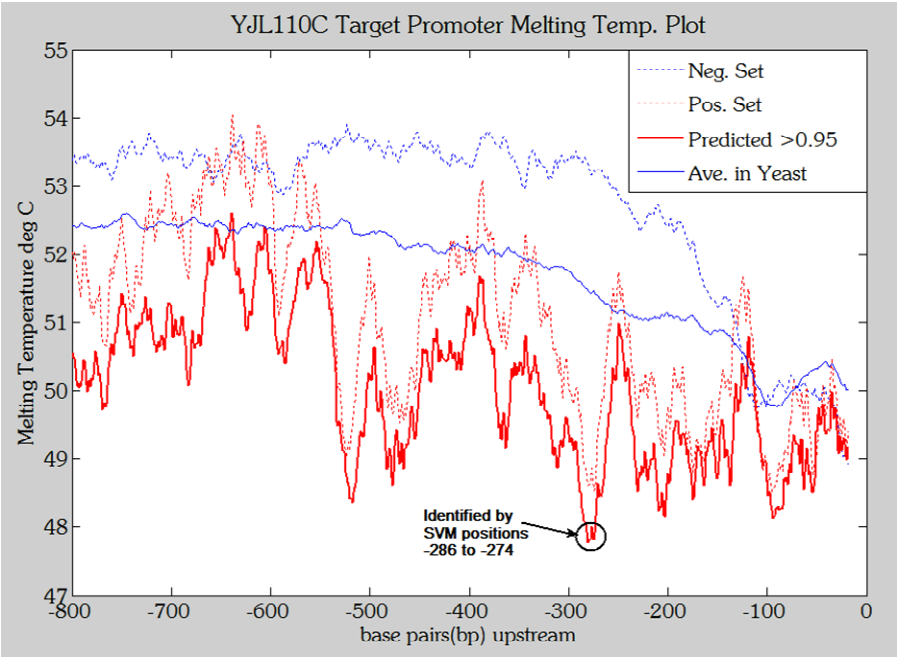


Figure 8 - YJL110C Target Melting Temperature Plot

Using a 20bp window for DNA melting temperature calculation, temperature plots are presented for the average over all 5571 yeast genes (solid blue), positive targets for YJL110C (dashed red), negatives for YJL110C (dashed blue), and high confidence targets (solid red—P(true|distance to separator) ≥ 0.95) determined using Platt's method for probability assignment to SVM output. Targets of YJL110C have a lower melting temperature than the average or negative set gene. SVM feature ranking correctly identifies the window positions in which the target melting temperature is most unlike that for negative genes, suggesting altered promoter structure of the targets, which is conserved at positions -286 to -274.

The classifier for Ash1 has a prediction accuracy of 88%. The highest ranked feature is expression in She4 deletion mutants. Expression of known targets and new predictions is significantly different than in the negative set ($p=5.4e^{-15}$ and $p=6.5e^{-8}$

respectively). Ash1 mRNA is localized to the daughter cell late in the cell cycle, where it prevents mating type switching by repressing HO genes[35, 36]. Localization of Ash1 mRNA is dependent on She genes. Specifically She4 binds to the myosin motor protein She1, and deletion of She4 causes loss of actin polarization[35, 37]. Thus, deletion of She4 prevents appropriate localization and regulation by Ash1, resulting in the observed increased expression of Ash1 target genes.

Conclusions

Transcription factor binding site prediction is a difficult problem in computational biology. Our SVM-based approach generates classifiers for each TF, and the reliability of the predictors is assessed using cross validation. The selection of high confidence predictions is made simpler by the calculation of a posterior probability for each potential target gene. By incorporating many types of genome-wide measurements into a robust feature ranking system, it is possible to discover important biological aspects of regulation which are specific to each TF being studied. This has been demonstrated on the yeast cell cycle regulator, Swi6. The predicted targets of Swi6 match the known biology of the regulator and suggest possible new roles of action in bioactive lipid metabolism and the DNA damage response. Moreover, feature ranking has identified interesting biological properties of the regulator including expression change of its targets in Mbp1 deletion mutants, and over-representation/conservation of the motif ACGCG. Similar analyses can be carried out with other TFs, as shown with Gzf3 and Ash1, for which meaningful biological features are identified. Investigators may download predictions made for all TFs, view classifier accuracies, and download lists of top-ranked features for each regulator at the provided web server. Custom analyses of the full yeast transcriptional network can also be accessed online in the VisAnt browser.

The next step of this analysis is to apply these methods to the human genome and assess their reliability. The possibility exists for the development of an electronic-chip-ChIP in human genome, whereby thousands of predictions can be made and the most reliable of these can be tested experimentally.

Methods

SVM training and validation

The SVM is a statistical learning method originally developed by Vapnik [38]. SVMs are based on rigorous statistical principles and show excellent performance when making predictions on many types of large genomic datasets. The algorithm seeks a maximal separation between two groups of binary labeled (e.g., 0, 1 or negative, positive) training examples [39]. The training examples are feature vectors **x** of individual genes, each vector populated by measurements taken on genome scale datasets (see below). These measurements are the attributes of the data. A single classifier based on these features is then constructed to predict targets for each TF. Positives are genes which are known targets of the TF, and negatives are a randomly chosen subset of genes (equal in size to the positive set) which are least likely to be targets. The separation of targets from non-targets is accomplished by an optimization which finds a hyperplane separating the two classes. The hyperplane is chosen to be as distant as possible from the data points, thus creating a *maximal-margin* hyperplane. The classifier can then be applied to new, unlabeled genes. We

have successfully applied SVMs to regulatory predictions before[17, 18, 40], and an in-depth tutorial on SVM training is available on our website.

The *C* parameter must be specified in SVM training. This parameter adjusts the tolerance of the algorithm for misclassifications. As with feature selection (described below), the classifier for YIR018W was used as the prototype for parameter selection. Grid selection was performed on the training set for YIR018W using many values of *C*, and classifier accuracy was measured with 5-fold cross validation. The SVM was seen to be insensitive to the choice of *C*, with most values less than 1 showing similar performance. Tested values include: [2^{-7} 2^{-5} 2^{-3} 2^{-1} 1 1.5 2 2^2 2^3 2^4 2^5 2^6]. The value 0.0078 was chosen as this was the value reported by the SPIDER machine learning package[41] as having the best performance.

Choosing negatives for TF target prediction can be difficult, since there is no defined set of genes known *not* to be targets. As in our previous work, ChIP-chip results can serve as a guide. For every TF, ChIP-chip results are used to identify genes which have the highest *p*-values (least significant) for binding under all tested conditions. The number of negatives is chosen to be at least three times the size of the positive set, or at least 600 genes, whichever is larger. Classifiers constructed on different randomly chosen negatives may give different results, since some unknown targets may be incorrectly assigned to the negative set. To smooth out these fluctuations, 100 classifiers are constructed for each transcription factor using a random resampling from the negative set. Each resampling is equal in size to the positive set and all 100 classifiers are tested using leave-one-out cross validation (LOOCV). The final performance statistics (accuracy, PPV, etc.) are averages from the 100 trials. The scheme used here for classifier construction is outlined in figure 1. To illustrate the construction and validation more concretely, a short outline is provided below. For an example TF-A:

1. Assemble positive set. Sample *n* genes randomly from the negative pool (see above) to construct the negative set (*n* = number of known targets).
2. Spit the data for LOOCV.
3. Use SVM-RFE to rank all features in the training set.
4. Construct SVM classifier on top 1500 features. Save full feature ranking.
5. Classify left out gene.
6. Repeat steps 2-5 to complete LOOCV. Save all feature rankings.
7. Calculate performance statistics (Accuracy, PPV, etc.)
8. Repeat steps 1-7 100 times.
9. Calculate final performance statistics (i.e., mean Accuracy, mean PPV, etc.).

A new gene can be classified by applying all 100 classifiers for TF-A to the feature vector for that gene. Each classification produces a posterior probability (see below), and the mean of all 100 probabilities is calculated. If $P>0.5$, classify the gene as a target of TF-A. The full set of feature rakings on every training set is used to calculate the final feature rank (see below).

Classifying new targets and prediction significance

As described in [19], SVMs can provide a probabilistic output which in this case measures the likelihood that any given gene is a target. This is given in the form of a class conditional probability, *P*(target | SVM output), where "SVM output" is the distance of the gene from the separating hyperplane. These outputs can be referred to as Platt's posterior probabilities (after the author of [19]) or simply as the true positive probability. The intuition of this method of assigning probabilities is that data points

which are deeper in the positive region (i.e., further from negative examples) are the most likely to be true positives. Our classifiers are trained for probabilistic output as recommended in [19], and new genes are classified using the average probability assigned by all 100 classifiers for a given TF. An average posterior probability greater then 0.5 is considered to yield a positive. Throughout the manuscript we refer to this probabilistic output using the upper-case *P* (e.g., $P>0.99$), whereas *p*-values measured by other means are shown in lower-case (e.g., $p<0.01$).

Genomic feature selection and ranking

The SVM algorithm can be used to select and rank data features. An important output from the algorithm is the vector **w**, which contains the learned weights of each feature. This vector points in a direction perpendicular to the hyperplane, and thus defines its orientation. Features with higher components in **w** are more useful in separating the positive and negative classes. The SVM recursive feature elimination (SVM-RFE) algorithm uses **w** to select features useful for classification[42]. The original SVM-RFE algorithm trains an SVM on a training set, and the components (attributes) of the feature vector **x** which have smallest weights are discarded [42]. The process is repeated until the desired number of attributes remains. In our study, half of the features are removed during each iteration until 1550 features remain. Features are then removed one at a time until the target of 1500 is reached. As indicated in the Discussion, the target of 1500 is determined by exploring the effect of feature selection on the prototype TF-classifier for YIR018W.

After this feature selection, the **w**-vector for the top 1500 features is used to determine the rank of the features in that training set, with higher weighted features having higher rank. These rankings are accumulated over every training set during cross validation of all 100 classifiers created for a TF. The result is a large set of feature rankings for a particular factor. The top 40 features from each ranking are collected into a list, and a count is taken of the number of times each feature appears. The final rank is established by sorting the features based on the frequency of their appearance. Therefore, features which are consistently ranked high during all cross-validation trials are given a high rank. Clearly, features high on this list are reliably important for separation and robust to changes in the training set.

Feature Datasets

Eight different types of features were used to describe genes. The first six feature sets have been used previously and their full descriptions can be found in [18]. The remaining three datasets have been modified or are novel.

1. *k*-mers (K-MER)—The distribution of all *k*-mers in a gene's promoter may be used to predict whether it is bound or not-bound by a TF. Feature vectors are formed by enumerating all possible strings of nucleotides of length 4, 5, and 6. The number of occurrences of each string is counted in a gene's promoter region, and this string of counts is the feature vector for the gene.
2. *k*-mers with Mismatch (M01)—Similar to *k*-mer counts, occurrences of all strings of length 4, 5, and 6 are counted. In addition, any string which contains only one mismatch is also considered a hit, but is given a count of 0.1 rather than 1.
3. Melting Temperature Profile (MT)—It is possible that TF binding is facilitated by conformational adjustments in promoter DNA, which depends on the stability of the helix. Some recent evidence shows correlation between sites of promoter melting, regulatory sites, and transcription initiation sites[43]. Our previous analysis also

demonstrates that the profiles of melting temperatures along promoters are significantly different for target and non-target genes for some TFs[18]. The EMBOSS[44] toolbox is used to calculate the melting temperature profiles of all yeast promoters using a sliding window of 20bp. The feature vectors are the same as described in [18].

4. Homolog Conservation (HC)—[45] BLASTP is used to compare proteins in yeast to those in 180 prokaryotic genomes. The best hit *E*-values to each genome are discretized by placing them into one of six bins using empirically determined *E*-value cut-offs. Bin numbers range from 0 (no significant hit) to 5 (very significant). Each gene then has 180 features, each for a different genome, with values ranging from 0-5, signifying the strength of the best BLASTP hit of that gene's protein to another genome.

6. *k*-mer Median Positions (Kpo)—For each possible *k*-mer (k = 4, 5, and 6) we record its median distance from the transcription start in each gene. If the transcription factor shows positional bias in promoter binding this dataset could be useful in generating a classifier.

7. Expression (EXP)—Normalized log2 ratios for each gene across 1011 experiments[46] are used as features. Each gene's expression profile is normalized to a mean of 0 and standard deviation of 1. For each gene a vector 1011 long (one feature for each expression experiment) is included in the data set.

8. *k*-mer Overrepresentation (Kev)—This method counts the number of each *k*-mer appearing in a promoter and calculates the significance of its occurrence. This method is the same as that reported in our previous work[18], except that the binomial distribution is used to calculate *p*-values rather than the Poisson distribution. This is in line with the recommendations in RSA tools[47, 48]. Furthermore, instead of directly using *p*-values, *E*-values were calculated according to

$$Evalue = -\log_{10}\left(pvalue \times D\right),$$

where D is the number of *k*-mers in the analysis. *E*-values account for the fact that many *k*-mers are being analyzed, and their use is equivalent to a correction for multiple hypothesis testing. Higher *E*-values correspond to more significant *k*-mers.

8. Conserved *k*-mers—This method for constructing a *k*-mer conservation matrix is based on output generated by the PhastCons algorithm[49, 50]. PhastCons is a two state phylogenetic hidden Markov model. The underlying idea is that conserved elements evolve more slowly than non-conserved elements. Thus, it has a "slow" state for conserved DNA and a "fast" state for non-conserved, more rapidly changing sites. Given DNA sequence alignments from multiple species, PhastCons outputs a probability score for each base pair in the alignment indicating from which state the sequence arises. This probability can be interpreted as the likelihood that the base pair is part of a conserved element. Genomic alignments for seven yeast species are used to generate the probability scores, which are available for download from the USC genome browser website[51, 52].

During *k*-mer counting, each *k*-mer is given a unique weight depending on the average PhastCons score of its nucleotide positions. Simply weighting by the probabilities would result in missing data, since some genomic regions have no alignments. Instead we introduce a weighting scheme which increases the weight of a *k*-mer according to its conservation. If a *k*-mer is not conserved, it will simply receive a count of 1 as usual. Our weighting metric is:

$$\frac{1}{1-\beta P_c}$$

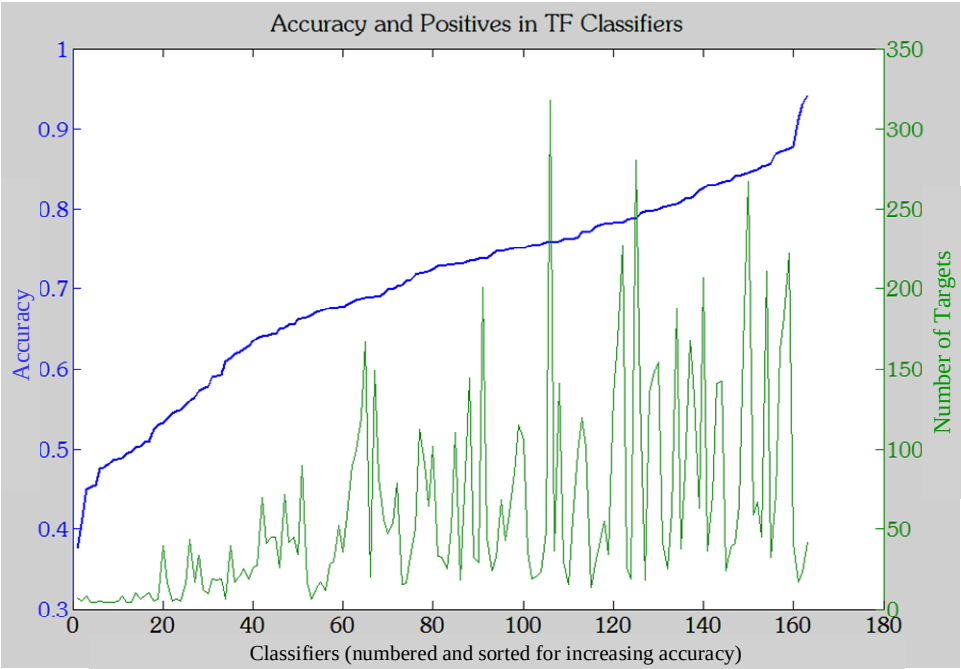
where P_c is the average PhastCons score for a particular *k*-mer. β is an adjustable parameter which controls how much the conservation of a *k*-mer increases its count. In this study we choose β = 0.75, so that an element with a maximum conservation of 1 has a count of 4. An element which shows no conservation has the default count of 1. Increasing β will further emphasize the effect of conservation. This method based on PhastCons is inspired by the "marginalized motif kernel for phylogenetic shadowing" introduced in [53]. Their method uses promoter alignments and a probabilistic model of fast and slow evolution to assess conserved elements. While their method can be considered more robust when good sequence alignments are available, we adopt the approach described here so that all yeast sequences may be included in our analysis. This will also be useful in the near future when we apply our method to the human genome, for which high quality alignments in intergenic regions are more sparse.

Functional Analysis

Statistical enrichment of GO biological process terms in gene sets was performed using the GO Term Finder on the Saccharomyces Genome Database website[54].

VisAnt Networks

The networks (such as Figure3) created with the VisAnt toolkit show links which have come from many publications. Any particular type of link (e.g., protein-protein interaction) may represent a collection of data from several genomic datasets. Each link type is referred to in VisAnt as a "method" and each method has a unique identifier. The method IDs for the link types in this paper include: M0037(phylogenetic profile), M0013(copurification), M0040(screened yeast-2-hybrid), M0031(other biophysical), M0046(Bayesian Predicted Interaction), M0045(affinity chromatography). Complete references and datasets are available in the VisAnt suite, accessible from the website http://visant.bu.edu/.



Supplementary Figure 1– Accuracy and Number of Positives
This figure plots the classifier accuracy on the left-y-axis(blue), and the number of positives (targets) on the right-y-axis(green). Classifiers are numbered on the x-axis and sorted according to increasing accuracy. A loose trend is present, showing that increasing the number of positives increases classifier accuracy. This is mainly seen when 50 or fewer positives exist. Classifiers with 20 or fewer examples tend to do poorly.

GOID	GO_term	Frequency	Genome F	Probability	Gene(s)	Directly An
722	telomerase	6 out of 60	14 out of 7	2.29E-09	YRF1-1 YF	722
6312	mitotic rec	6 out of 60	26 out of 7	8.72E-08	YRF1-1 YF	722
723	telomere m	7 out of 60	180 out of	0.00068	YRF1-1 YF	722:723
6310	DNA recor	7 out of 60	192 out of	0.00099	YRF1-1 YF	722:6319
7001	chromosor	8 out of 60	374 out of	0.0113	YRF1-1 YF	722:1308:7
51276	chromosor	8 out of 60	385 out of	0.01328	YRF1-1 YF	722:1308:7
45941	positive re	3 out of 60	69 out of 7	0.01942	SFG1 UTH	45944:130
45935	positive re	3 out of 60	70 out of 7	0.02016	SFG1 UTH	45944:130
9893	positive re	3 out of 60	72 out of 7	0.02169	SFG1 UTH	45944:130
31325	positive re	3 out of 60	72 out of 7	0.02169	SFG1 UTH	45944:130
15980	energy der	5 out of 60	201 out of	0.02483	QCR6 GID	9060:4572
48522	positive re	3 out of 60	76 out of 7	0.02493	SFG1 UTH	45944:130
51242	positive re	3 out of 60	76 out of 7	0.02493	SFG1 UTH	45944:130
43119	positive re	3 out of 60	76 out of 7	0.02493	SFG1 UTH	45944:130
6418	tRNA amin	2 out of 60	30 out of 7	0.02558	MSR1 FRS	6420:6432
43038	amino acid	2 out of 60	30 out of 7	0.02558	MSR1 FRS	6420:6432
43039	tRNA amin	2 out of 60	30 out of 7	0.02558	MSR1 FRS	6420:6432
44262	cellular ca	5 out of 60	204 out of	0.02625	GID8 TSL1	45721:599
48518	positive re	3 out of 60	84 out of 7	0.03216	SFG1 UTH	45944:130
6139	nucleobas	19 out of 6	1526 out o	0.03405	YRF1-1 YF	722:45944
45229	external en	4 out of 60	150 out of	0.03514	SED1 EXC	7047
7047	cell wall or	4 out of 60	150 out of	0.03514	SED1 EXC	7047
6112	energy res	2 out of 60	36 out of 7	0.03571	TSL1 BMH	5992:5977
5975	carbohydr	5 out of 60	223 out of	0.03645	GID8 TSL1	45721:599
6259	DNA metal	9 out of 60	564 out of	0.0403	YRF1-1 YF	722:1308:6
6091	generation	5 out of 60	233 out of	0.04269	QCR6 GID	9060:6122
6066	alcohol me	4 out of 60	161 out of	0.04367	ERG26 GI	6696:4572
6073	glucan me	2 out of 60	41 out of 7	0.04513	EXG1 BMH	6073:5977
7124	pseudohyp	2 out of 60	50 out of 7	0.06408	SFG1 BMH	7124
5976	polysaccha	2 out of 60	56 out of 7	0.07794	EXG1 BMH	6073:5977
44264	cellular pol	2 out of 60	56 out of 7	0.07794	EXG1 BMH	6073:5977
45893	positive re	2 out of 60	64 out of 7	0.09772	SFG1 UTH	45944:130

Supplementary Table 1 – Significant Functions of Highly Regulated Genes
This file is the output from the GO Term Finder at the Saccharomyces Genome database. Using only classifiers which had an accuracy of 0.6 and targets identified with a posterior probability ≥ 0.95, the genes regulated by 12 or more TFs were input into the GO Term Finder. The results show statistically enriched GO terms, p-values, and provide the genes annotated to those terms.

GOID	GO_term	Frequency	Genome F	Probability	Gene(s)	Directly Ar
6350	transcriptio	11 out of 1	493 out of	9.25E-12	ABF1 RCS	45944:304
6355	regulation	9 out of 13	311 out of	2.85E-10	ABF1 RCS	45944:304
45449	regulation	9 out of 13	331 out of	4.95E-10	ABF1 RCS	45944:304
19219	regulation	9 out of 13	362 out of	1.09E-09	ABF1 RCS	45944:304
6139	nucleobas	13 out of 1	1526 out o	1.47E-09	ABF1 RCS	45944:304
31323	regulation	9 out of 13	419 out of	3.94E-09	ABF1 RCS	45944:304
51244	regulation	10 out of 1	613 out of	3.97E-09	ABF1 RCS	45944:304
50794	regulation	10 out of 1	614 out of	4.03E-09	ABF1 RCS	45944:304
50791	regulation	10 out of 1	631 out of	5.26E-09	ABF1 RCS	45944:304
50789	regulation	10 out of 1	641 out of	6.13E-09	ABF1 RCS	45944:304
19222	regulation	9 out of 13	445 out of	6.69E-09	ABF1 RCS	45944:304
6351	transcriptio	9 out of 13	450 out of	7.38E-09	ABF1 RCS	45944:304
6357	regulation	7 out of 13	187 out of	1.09E-08	ABF1 RCS	45944:122
6366	transcriptio	7 out of 13	285 out of	1.94E-07	ABF1 RCS	45944:122
45944	positive re	4 out of 13	53 out of 7	1.89E-06	ABF1 RCS	45944:732
45893	positive re	4 out of 13	64 out of 7	3.98E-06	ABF1 RCS	45944:732
45941	positive re	4 out of 13	69 out of 7	5.35E-06	ABF1 RCS	45944:732
45935	positive re	4 out of 13	70 out of 7	5.66E-06	ABF1 RCS	45944:732
9893	positive re	4 out of 13	72 out of 7	6.32E-06	ABF1 RCS	45944:732
31325	positive re	4 out of 13	72 out of 7	6.32E-06	ABF1 RCS	45944:732
43119	positive re	4 out of 13	76 out of 7	7.82E-06	ABF1 RCS	45944:732
48522	positive re	4 out of 13	76 out of 7	7.82E-06	ABF1 RCS	45944:732
51242	positive re	4 out of 13	76 out of 7	7.82E-06	ABF1 RCS	45944:732
48518	positive re	4 out of 13	84 out of 7	1.15E-05	ABF1 RCS	45944:732
44238	primary m	13 out of 1	3206 out o	2.29E-05	ABF1 RCS	45944:304
44237	cellular me	13 out of 1	3400 out o	4.92E-05	ABF1 RCS	45944:304
8152	metabolisn	13 out of 1	3490 out o	6.91E-05	ABF1 RCS	45944:304
9628	response t	5 out of 13	314 out of	0.00014	SKN7 NRG	6979:6970
50896	response t	6 out of 13	588 out of	0.00028	ABF1 SKN	715:6979:6
42221	response t	4 out of 13	233 out of	0.00059	SKN7 MSN	6979:1324
1403	invasive gr	2 out of 13	34 out of 7	0.00163	NRG1 STE	1403
50875	cellular phy	13 out of 1	4722 out o	0.00352	ABF1 RCS	45944:304
9987	cellular pro	13 out of 1	4761 out o	0.00391	ABF1 RCS	45944:304
6979	response t	2 out of 13	54 out of 7	0.00405	SKN7 MSN	6979:1324
7582	physiologic	13 out of 1	4790 out o	0.00423	ABF1 RCS	45944:304
6800	oxygen an	2 out of 13	56 out of 7	0.00434	SKN7 MSN	6979:1324
6970	response t	2 out of 13	61 out of 7	0.00513	SKN7 CIN	6970:9651
6950	response t	4 out of 13	426 out of	0.00543	ABF1 SKN	715:6979:6
30447	filamentou	2 out of 13	91 out of 7	0.01108	NRG1 STE	1403:7124
6333	chromatin	2 out of 13	100 out of	0.01326	ABF1 CBF	30466:633
6259	DNA metal	4 out of 13	564 out of	0.01449	ABF1 CBF	30466:715
6260	DNA replic	2 out of 13	105 out of	0.01455	ABF1 MBF	6260
45892	negative re	2 out of 13	125 out of	0.02021	ABF1 UME	30466:122
40007	growth	2 out of 13	125 out of	0.02021	NRG1 STE	1403:7124
16481	negative re	2 out of 13	129 out of	0.02144	ABF1 UME	30466:122
45934	negative re	2 out of 13	139 out of	0.02465	ABF1 UME	30466:122
7001	chromosor	3 out of 13	374 out of	0.02619	ABF1 CBF	30466:633
51276	chromosor	3 out of 13	385 out of	0.02824	ABF1 CBF	30466:633
31324	negative re	2 out of 13	151 out of	0.02874	ABF1 UME	30466:122
9892	negative re	2 out of 13	158 out of	0.03125	ABF1 UME	30466:122
7049	cell cycle	3 out of 13	406 out of	0.03241	MBP1 UMI	74:45836:7
51243	negative re	2 out of 13	171 out of	0.03613	ABF1 UME	30466:122
48523	negative re	2 out of 13	171 out of	0.03613	ABF1 UME	30466:122
43118	negative re	2 out of 13	173 out of	0.0369	ABF1 UME	30466:122
48519	negative re	2 out of 13	179 out of	0.03927	ABF1 UME	30466:122
50876	reproducti	2 out of 13	197 out of	0.04672	STE12 UM	747:30437
48610	reproducti	2 out of 13	197 out of	0.04672	STE12 UM	747:30437
6325	establishm	2 out of 13	211 out of	0.05285	ABF1 CBF	30466:633
6323	DNA packa	2 out of 13	211 out of	0.05285	ABF1 CBF	30466:633
278	mitotic cell	2 out of 13	235 out of	0.064	UME6 SW	7068:82
3	reproducti	2 out of 13	263 out of	0.07796	STE12 UM	747:30437

Supplementary Table 2 – Significant Function of Master Regulators
This file is the output from the GO Term Finder at the Saccharomyces Genome database. Using only classifiers which had an accuracy of 0.6 and targets identified with a posterior probability ≥0.95, the regulators which are predicted to bind to 300 or more genes were used as input to the GO Term Finder.

Authors' contributions

DH coded the required software in Matlab and Perl, conceived of many of the design implementations, and wrote this article. All authors made contributions to this manuscript and the experimental design. CD initially conceived and motivated this work. All authors read and approved the final manuscript.

References

1. Z Hu, J Mellor, J Wu, C DeLisi: **VisANT: an online visualization and analysis tool for biological interaction data**. *BMC Bioinformatics* 2004, **5**:17.
2. Z Hu, J Mellor, J Wu, T Yamada, D Holloway, C DeLisi: **VisANT: data-integrating visual framework for biological networks and modules**. *Nucl. Acids Res.* 2005, **33**:W352-357.
3. GD Stormo: **DNA Binding Sites: Representation and Discovery.** *Bioinformatics* 2000, **16**:16-23.
4. CT Workman, GD Stormo: **ANN-Spec: a method for discovering transcription factor binding sites with improved specificity**. *Pac Symp Biocomput* 2000:467-78.
5. TD Schneider, GD Stormo, L Gold, A Ehrenfeucht: **Information content of binding sites on nucleotide sequences**. *Journal of Molecular Biology* 1986, **188**:415-431.
6. T Schneider, R Stephens: **Sequence logos: a new way to display consensus sequences**. *Nucl. Acids Res.* 1990, **18**:6097-6100.
7. EM Conlon, XS Liu, JD Lieb, JS Liu: **Integrating regulatory motif discovery and genome-wide expression analysis**. *PNAS* 2003, **100**:3339-3344.
8. S Keles, MJ van der Laan, C Vulpe: **Regulatory motif finding by logic regression**. *Bioinformatics* 2004, **20**:2799-2811.
9. W Wang, JM Cherry, D Botstein, H Li: **A systematic approach to reconstructing transcription networks in *Saccharomyces scerevisiae***. *PNAS* 2002, **99**:16893-16898.
10. H Bussemaker, H Li, E Siggia: **Regulatory Element Detection Using Correlation with Expression**. *Nature Genetics* 2001, **27**:167-171.
11. K Birnbaum, PN Benfey, DE Shasha: **cis Element/Transcription Factor Analysis (cis/TF): A Method for Discovering Transcription Factor/cis Element Relationships**. *Genome Res.* 2001, **11**:1567-1573.
12. Z Zhu, Y Pilpel, G Church: **Computational Identification of Transcription Factor Binding Sites via a Transcription-Factor-Centric-Clustering (TFCC) Algorithm**. *Journal of Molecular Biology* 2002, **318**:71-81.

13. M Pritsker, Y-C Liu, MA Beer, S Tavazoie: **Whole-Genome Discovery of Transcription Factor Binding Sites by Network-Level Conservation**. *Genome Res.* 2004, **14**:99-108.
14. S Elemento, S Tavazoie: **Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach**. *Genome Biology* 2005, **6**:R18.
15. M Tompa, et al.: **Assessing computational tools for the discovery of transcription factor binding sites**. *Nature Biotechnology* 2005, **23**:137-144.
16. JW Fickett: **Coordinate Positioning of MEF2 and Myogenin Binding Sites**. *Gene* 1996, **172**:19-32.
17. D Holloway, M Kon, C DeLisi: **Machine Learning Methods for Transcription Data Integration: in press**. *IBM Journal of Research and Development on Systems Biology* 2006.
18. D Holloway, M Kon, C DeLisi: **Machine Learning for Predicting Targets of Transcription Factors in Yeast: in press**. *Synthetic and Systems Biology* 2006.
19. JC Platt: **Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods**. *Advances in Large Margin Classifiers, MIT Press* 1999.
20. TI Lee, NJ Rinaldi, F Robert, DT Odom, Z Bar-Joseph, GK Gerber, NM Hannett, CT Harbison, CM Thompson, I Simon, et al: **Transcriptional Regulatory Networks in Saccharomyces cerevisiae**. *Science* 2002, **298**:799-804.
21. TI Lee, NJ Rinaldi, F Robert, DT Odom, Z Bar-Joseph, GK Gerber, NM Hannett, CT Harbison, CM Thompson, I Simon, et al: **Web site: [http://jura.wi.mit.edu/cgi-bin/young_public/navframe.cgi?s=17&f=downloaddata]**.
22. CT Harbison, DB Gordon, TI Lee, NJ Rinaldi, KD Macisaac, TW Danford, NM Hannett, J-B Tagne, DB Reynolds, J Yoo, et al: **Transcriptional regulatory code of a eukaryotic genome**. *Nature* 2004, **431**:99-104.
23. M Huang, Z Zhou, SJ Elledge: **The DNA Replication and Damage Checkpoint Pathways Induce Transcription by Inhibition of the Crt1 Repressor**. *Cell* 1998, **94**:595-605.
24. PT Spellman, G Sherlock, MQ Zhang, VR Iyer, K Anders, MB Eisen, PO Brown, D Botstein, B Futcher: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization**. *Mol. Biol. Cell* 1998, **9**:3273-3297.
25. JM Sidorova, LL Breeden: **Rad53-dependent phosphorylation of Swi6 and down-regulation of CLN1 and CLN2 transcription occur in response to DNA damage in Saccharomyces cerevisiae**. *Genes Dev.* 1997, **11**:3032-3045.
26. AP Gasch, M Huang, S Metzner, D Botstein, SJ Elledge, PO Brown: **Genomic Expression Responses to DNA-damaging Agents and the Regulatory Role of the Yeast ATR Homolog Mec1p**. *Mol. Biol. Cell* 2001, **12**:2987-3003.
27. G Giaever, AM Chu, L Ni, C Connelly, L Riles, S Veronneau, S Dow, A Lucau-Danila, K Anderson, B Andre, et al: **Functional profiling of the Saccharomyces cerevisiae genome**. *Nature* 2002, **418**:387-391.

28. CE Horak, NM Luscombe, J Qian, P Bertone, S Piccirrillo, M Gerstein, M Snyder: **Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae**. *Genes Dev.* 2002, **16**:3017-3033.
29. S Leem, C Chung, Y Sunwoo, H Araki: **Meiotic role of SWI6 in Saccharomyces cerevisiae**. *Nucl. Acids Res.* 1998, **26**:3154-3158.
30. V Matys, OV Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, et al: **TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes**. *Nucl. Acids Res.* 2006, **34**:D108-110.
31. SV de Avalos, Y Okamoto, YA Hannun: **Activation and Localization of Inositol Phosphosphingolipid Phospholipase C, Isc1p, to the Mitochondria during Growth of Saccharomyces cerevisiae**. *J. Biol. Chem.* 2004, **279**:11537-11545.
32. Y Okamoto, SV de Avalos, Y Hannun: **Functional Analysis of ISC1 by Site-Directed Mutagenesis**. *Biochemistry* 2003, **42**:7855-7862.
33. L Cowart, Y Okamoto, X Lu, Y Hannun: **Distinct roles for de novo versus hydrolytic pathways of sphingolipid biosynthesis in Saccharomyces cerevisiae**. *Biochemical Journal* 2006, **393**:733-740.
34. PC Reynolds, BJ Maurer, RN Kolesnick: **Ceramide synthesis and metabolism as a target for cancer therapy**. *Cancer Letters* 2004, **206**:169-180.
35. MP Cosma: **Daughter-specific repression of Saccharomyces cerevisiae HO: Ash1 is the commander**. *EMBO reports* 2005, **5**:953-957.
36. H Toi, K Fujimura-Kamada, K Irie, Y Takai, S Todo, K Tanaka: **She4p/Dim1p Interacts with the Motor Domain of Unconventional Myosins in the Budding Yeast, Saccharomyces cerevisiae**. *Mol. Biol. Cell* 2003, **14**:2237-2249.
37. B Wendland, J McCaffery, Q Xiao, S Emr: **A novel fluorescence-activated cell sorter-based screen for yeast endocytosis mutants identifies a yeast homologue of mammalian**. *J. Cell Biol.* 1996, **135**:1485-1500.
38. V Vapnik: **Statistical Learning Theory**. *Text:The Nature of Statistical Learning Theory* 1998.
39. B Sholkopf, AJ Smola: **Learning with Kernels**. *MIT Press* 2002.
40. D Holloway, M Kon, C DeLisi: **Integrating genomic data to predict transcription factor binding**. *Proc. of the Workshop on Genome Informatics* 2005, **16**:83-94.
41. J Weston, A Elisseeff, G Bakir, F Sinz, et al: **SPIDER: object oriented machine learning library: [http://www.kyb.tuebingen.mpg.de/bs/people/spider/]**.
42. I Guyon, J Weston, S Barnhill, V Vapnik: **Gene Selection for Cancer Classification using Support Vector Machines**. *Machine Learning* 2002, **46**:389-422.
43. CH Choi, G Kalosakas, KO Rasmussen, M Hiromura, AR Bishop, A Usheva: **DNA dynamically directs its own transcription initiation**. *Nucl. Acids Res.* 2004, **32**:1584-1590.
44. P Rice, I Longden, A Bleasby: **EMBOSS: The European Molecular Biology Open Software Suite**. *Trends in Genetics* 2000, **16**:276-277.
45. E Snitkin, A Gustafson, C DeLisi: *Unpublished work Personal Communication*.

46. S Bergman, J Ihmels, N Barkai: **Iterative Signature Algorithm for the Analysis of Large-Scale Gene Expression Data**. *Physical Review* 2003, **67**.
47. J van Helden, J Collado-Vides: **Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies**. *Journal of Molecular Biology* 1998, **281**:827-842.
48. J van Helden: **Regulatory sequence analysis tools**. *Nucleic Acids Research* 2003, **31**:3593-3596.
49. A Siepel, D Haussler: **Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis**. *Journal of Computational Biology* 2004, **11**:413-428.
50. A Siepel, G Bejerano, JS Pedersen, AS Hinrichs, M Hou, K Rosenbloom, H Clawson, J Spieth, LW Hillier, S Richards, et al: **Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes**. *Genome Res.* 2005, **15**:1034-1050.
51. D Karolchik, R Baertsch, M Diekhans, TS Furey, A Hinrichs, YT Lu, KM Roskin, M Schwartz, CW Sugnet, DJ Thomas, et al: **The UCSC Genome Browser Database**. *Nucl. Acids Res.* 2003, **31**:51-54.
52. D Karolchik, AS Hinrichs, TS Furey, KM Roskin, CW Sugnet, D Haussler, WJ Kent: **The UCSC Table Browser data retrieval tool**. *Nucl. Acids Res.* 2004, **32**:D493-496.
53. J-P Vert, R Thurman, WS Noble: **Kernels for Gene Regulatory Regions**. *Proceedings of the 19th Annual Conference on Neural and Information Systems, Vancouver, BC* 2005.
54. E Hong, R Balakrishnan, K Christie, M Costanzo, S Dwight, S Engel, F DG, J Hirschman, L MS, N R, et al: **Saccharomyces Genome Database: [http://www.yeastgenome.org/]**.