Integrating Genomic Data to Predict Transcription Factor Binding

Dustin T. Holloway[1] *Mark Kon*[2] *Charles DeLisi*[3]
dth128@bu.edu mkon@bu.edu delisi@bu.edu

[1] Molecular Biology Cell Biology and Biochemistry, Boston University, Boston, MA 02215, U.S.A.
[2] Department of Mathematics and Statistics, Boston University, Boston, MA 02215, U.S.A.
[3] Bioinformatics and Systems Biology, Boston University, Boston, MA 02215, U.S.A.

Abstract

Transcription factor binding sites (TFBS) in gene promoter regions are often predicted by using position specific scoring matrices (PSSMs), which summarize sequence patterns of experimentally determined TF binding sites. Although PSSMs are more reliable than simple consensus string matching in predicting a true binding site, they generally result in high numbers of false positive hits. This study attempts to reduce the number of false positive matches and generate new predictions by integrating various types of genomic data by two methods: a Bayesian allocation procedure, and support vector machine classification.

Several methods will be explored to strengthen the prediction of a true TFBS in the *Saccharomyces cerevisiae* genome: binding site degeneracy, binding site conservation, phylogenetic profiling, TF binding site clustering, gene expression profiles, GO functional annotation, and k-mer counts in promoter regions. Binding site degeneracy (or redundancy) refers to the number of times a particular transcription factor's binding motif is discovered in the upstream region of a gene. Phylogenetic conservation takes into account the number of orthologous upstream regions in other genomes that contain a particular binding site. Phylogenetic profiling refers to the presence or absence of a gene across a large set of genomes. Binding site clusters are statistically significant clusters of TF binding sites detected by the algorithm ClusterBuster. Gene expression takes into account the idea that when the gene expression profiles of a transcription factor and a potential target gene are correlated, then it is more likely that the gene is a genuine target. Also, genes with highly correlated expression profiles are often regulated by the same TF(s). The GO annotation data takes advantage of the idea that common transcription targets often have related function. Finally, the distribution of the counts of all k-mers of length 4, 5, and 6 in gene's promoter region were examined as means to predict TF binding. In each case the data are compared to known true positives taken from ChIP-chip data[1, 2], Transfac, and the Saccharomyces Genome Database.

First, degeneracy, conservation, expression, and binding site clusters were examined independently and in combination via Bayesian allocation. Then, binding sites were predicted with a support vector machine (SVM) using all methods alone and in combination. The SVM works best when all genomic data are combined, but can also identify which methods contribute the most to accurate classification. On average, a support vector machine can classify binding sites with high sensitivity and an accuracy of almost 80%.

Keywords: Transcription Factor, Support Vector Machine, TFBS, binding site, motif.

1 Introduction

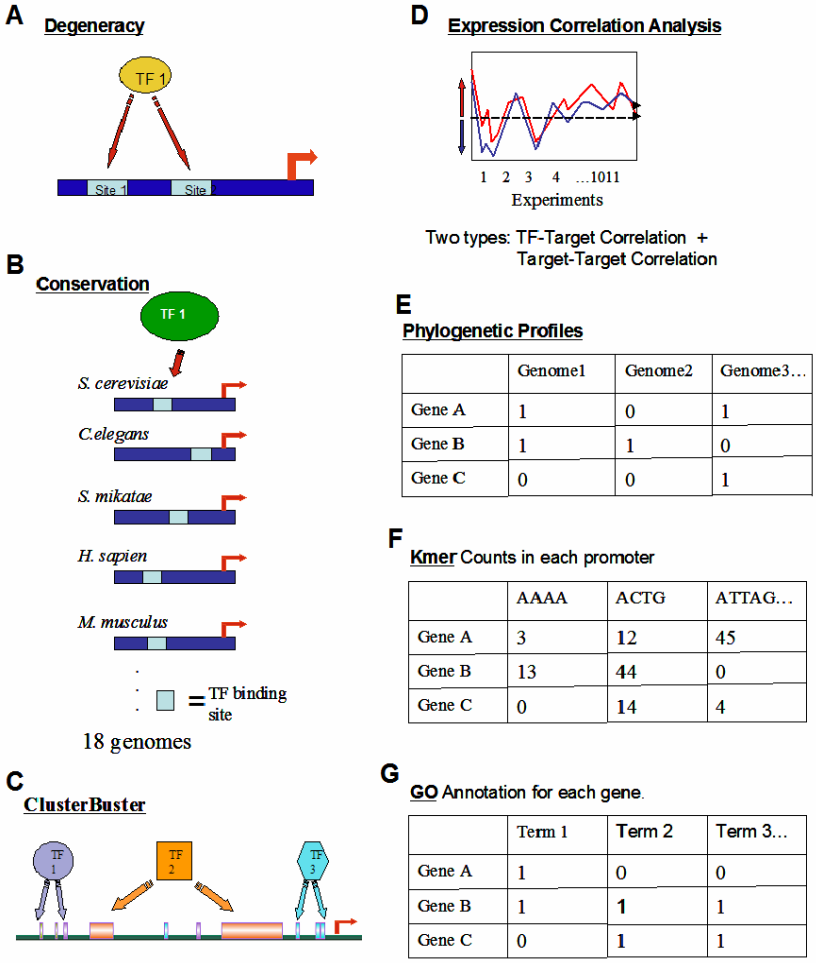
Transcriptional mechanisms are shaped at their most basic level by the direct interactions of transcription factors and the specific cis-elements that they bind in DNA. Understanding such interactions can provide insight into how cells respond to changing needs or stresses. Experimentally determined TF binding sites are often represented as consensus sequences or probability matrices. Although consensus strings are easy to interpret visually, they lack a quantitative description of possible nucleotide frequencies at each position in the binding site. A probability matrix (or position weight matrix) provides a detailed description of the target site and allows for the efficient scanning and comparison of any DNA sequence against the binding site model. A number of published algorithms are available to detect cis-elements using position probability matrices. The present analysis will find motifs using MotifScanner, which assumes a sequence model where regulatory elements are distributed within a noisy background sequence[3].

Until recently there have only been 17 detailed probability matrices available for *S. cerevisiae* despite the wealth of experimental binding site data for the estimated 203 TFs in yeast[4]. ChIP-chip experiments have assembled over 11,000 unique TF-gene interactions, and computational methods have been used to generate PWMs for many of these TFs[1, 5]. Combined with the matrices in Transfac and the literature, matrices for 104 transcription factors were available for this study.

Position specific scoring matrices are currently the most widely used method to represent TF binding preferences; however, binding models of this sort suffer from several disadvantages, namely that they assume a

finite binding site width, something which is not true for flexible TFs, and they assume that each nucleotide within a binding site is independent of the others. Despite their drawbacks, simple probability matrices are currently the standard in the field, being both easily available and easily used to make site predictions. Nevertheless, binding site detection is plagued by a high rate of false positive predictions, with some matrices producing predictions at a frequency of 1 in 500bp[6]. Because of this great excess of false positives, other information must be brought to bear in order to more accurately predict regulated genes.

All together, eight types of genomic data have been examined for their ability to predict transcription factor binding. The first is binding site degeneracy (I) (or TF motif redundancy) which explores the relationship between the number of predicted motifs for a TF in a promoter and the likelihood that the TF binds. More instances of a motif in a promoter region are expected to translate into a higher probability of real binding. Conservation (II) uses sequence information from multiple orthologous promoters to determine how often a predicted binding site occurs in other genomes. The more genomes in which this site is conserved, the more a motif sequence is likely to be a true binding site. Detection of clusters (III) of binding sites has also been shown to more accurately elucidate binding sites, and here we use ClusterBuster[7] to determine whether a predicted site lies in such a cluster. Gene expression profiles were used in two ways to add information to the analysis. TF-Target correlations (IV) will help predict instances when a transcription factor regulates genes via its own expression level, and target-target correlations (V) will identify genes having similar expression to known targets. Similarly, GO term annotation (VI) can find genes which have been annotated with very similar function to known targets, while phylogenetic profiles (VII) can identify those with a similar pattern of occurrence to known targets across 65 microbial genomes. Finally, the k-mer distribution (VIII) built from the counts of all k-mers of length 4, 5, and 6 in each gene's promoter can be used to predict whether any given gene has a distribution similar to known targets. Schematic descriptions of these methods can be seen in Figure 1. Again, known binding data from Transfac, ChIP-chip experiments, and the Saccharomyces Genome Database were used as a standard of comparison against which to measure predictions and train a support vector machine for each TF.



	Genome1	Genome2	Genome3…
Gene A	1	0	1
Gene B	1	1	0
Gene C	0	0	1

	AAAA	ACTG	ATTAG…
Gene A	3	12	45
Gene B	13	44	0
Gene C	0	14	4

	Term 1	Term 2	Term 3…
Gene A	1	0	0
Gene B	1	1	1
Gene C	0	1	1

Figure 1 Eight Genomic Data methods. **A**. Degeneracy is the frequency, or number, of TF motifs that appear upstream of a gene. **B.** Conservation measures the number of genomes in which a motif is conserved. **C** .ClusterBuster finds clusters of heterogeneous motifs from several TFs. Such clusters are more likely to contain real binding sites. **D** . Expression data comprises two methods 1. TF-Target correlations are measured, and 2. Target-Target correlations are discovered by comparing a gene's expression to that of known targets. **E**. Phylogenetic profiles show a gene's occurrence profile among many genomes. Common targets of a TF may share similar occurrence profiles **F.** K-mer counts in gene promoter regions can be used to differentiate targets from non-targets. **G.** GO term profiles of each gene are used as a measure of gene function. Functional similarity can potentially predict new targets of a TF since genes having common regulation are thought to share function.

Since each individual method is partially successful at predicting regulation, the consensus of many measures together will isolate interactions that are highly likely to be true. This is in fact what occurs when we take subsets of various datasets which show higher probability of containing binding sites (Bayesian allocation). But, although false positives are greatly reduced, sensitivity also decreases. Integrating the data using an SVM learning algorithm greatly enhanced our results by achieving high accuracy, reducing false predictions and reaching a sensitivity beyond sixty percent. Four methods were explored using Bayesian allocation (Degeneracy, Conservation, Clusters, TF-Target Correlation) while all methods were examined by support vector machine.

2 Methods and Results

All promoter sequences were collected from RSA tools, Ensembl, or the Broad Institute's Fungal Genome Anatomy Project[8-10]. Sequences were then masked [11, 12] where appropriate to exclude low complexity sequences and known repeat DNA from further analysis. The MotifScanner algorithm was used to scan all upstream regions for transcription factor binding sites using PSSM models. This algorithm requires a background sequence model, which in this case is a transition matrix of a 3rd order Markov model generated from the masked upstream regions of each genome. MotifScanner only requires one parameter be set by the user; this is the prior parameter, which can be interpreted as the probability that a given motif will be found by chance in a promoter. Several thresholds have been tested and the results reported in this paper are all at a setting of 0.15, which was found to be a reasonable middle ground in that it makes approximately 560 predictions per TF. Settings beyond 0.2 produced too many false hits to be useful.

The position specific scoring matrices (PSSMs) were adapted from Transfac *S. cerevisiae* count matrices (17 PSSMs) and from probability matrices published in[1] (107 PSSMs). Some PSSMs were redundant, and the cumulative set includes matrices for 104 TFs. All upstream regions from 18 eukaryotic genomes (ranging from human to yeast) were scanned with the assembled matrices. These matrices also served as input to ClusterBuster for predictions of binding site clusters.

2.1 Degeneracy as a Method to Predict Binding Sites

It is clear that a greater number of hits by a matrix in the upstream region of a gene will result in a greater likelihood that the TF for the matrix will actually bind the gene. For each prediction there is a certain probability that this prediction will be true, P(True|hit). It follows that if a certain upstream region has more than one hit, the probability that the factor binds is increased. This method hopes to better predict TF binding by taking into account the number of times the binding motif appears in a promoter. This method of degeneracy was used to evaluate the hits generated by weight matrices and the results show that repetition of TF motifs is directly proportional to the likelihood of finding a true binding site. This method's results are shown in Figure 2.

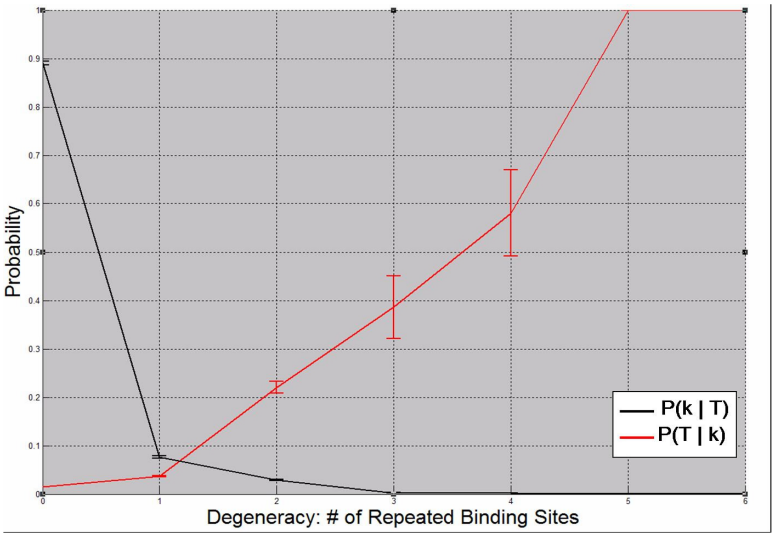


Figure 2 Degeneracy: Having more than one detected binding site for a TF in the upstream region of a gene increases the likelihood that the TF truly binds the gene. Higher counts of motifs yield fewer predictions; however, as the number of repetitions of a motif increases, the probability that the TF binds approaches 1. P(k|T)=Probability of motif count k given a set of True binding sites. P(T|k)=Probability of finding a True binding site given a motif count of k.

Data is average over 104 TFs.

2.2 Conservation as a Method to Predict Binding Sites

Comparative genomics tools have recently been applied with much success to the identification of transcription factor binding sites. Because most regulatory elements are in non-coding regions and show considerable variation in sequence even for the same TF, they aren't easily recognizable. However, binding sites are often preserved through evolution, and thus become apparent in what authors call a "footprint" in alignments of orthologous regions from different genomes. Cis-element conservation is a powerful way to detect functional non-

coding elements, and, in this case, will be modified and applied to 18 genomes ranging from yeast to human. Conservation of a TF binding site will be determined by counting hits of the TF probability matrix in orthologous upstream regions from several organisms. Orthology information was taken mainly from the Homologene database[13] for all organisms except for *sensu stricto* and *sensu lato* yeasts, which was downloaded from Washington University and the Whitehead Broad Institute [8, 14-16].

Previous studies have defined conservation as direct nucleotide conservation in *aligned* orthologous regions. In previous publications[14] this analysis involved manual inspection and modification of low scoring alignments, an approach that would be cumbersome and time consuming with a larger number of genomes. Other authors relied on whole genome alignments of closely related yeast species to identify orthologs and conserved upstream regions[15]. This strategy would be difficult if not impossible for genomes farther diverged than the few closely related yeast species. In this analysis, a hit by a PSSM in the upstream region of an ortholog counts as conservation. Some specificity will likely be lost with this strategy; however, the analysis should gain much power with the addition of so many more genomes. Also, since no alignment is made, conservation of a *potential binding site* is being measured rather than the exact nucleotide string. This is because a PSSM may identify sequences that are different in nucleotide composition but still match the probability matrix. This is a looser conservation criterion that makes sense biologically since natural selection will act to preserve a binding site, not necessarily an exact nucleotide string.

It is predicted that the stronger the conservation of a potential binding site, the more likely it is that the site is real one. This hypothesis is supported by results obtained (Figure 3). The probability of seeing a true site increases to 100% as binding site conservation reaches 15 genomes.

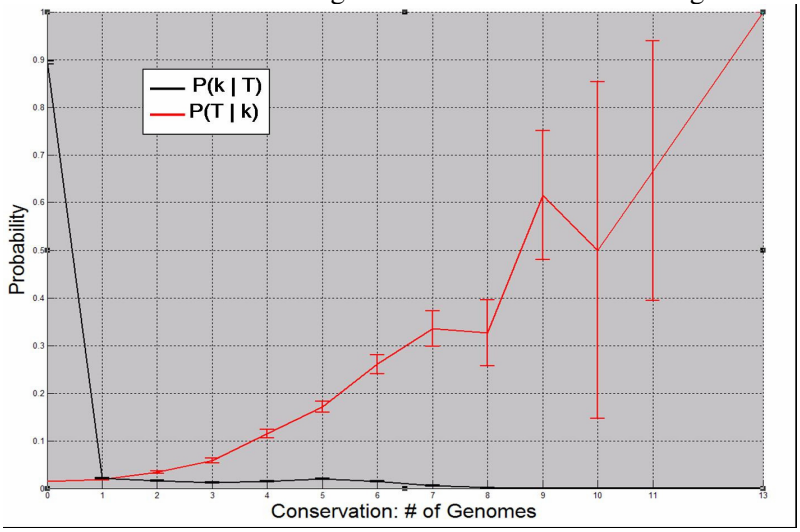


Figure 3 Conservation of a TF binding site in several orthologous upstream regions increases the likelihood that a potential site is a True site.

P(k|T)=Probability of site conservation in k genomes given a set of True binding sites.
P(T|k)=Probability of finding a True binding site given that it is conserved in k genomes.

Data is average over 104 TFs.

2.3 Motif Clusters as Indicators of Binding Sites

A third approach is to identify clusters of various TF binding sites. It is established that many transcription factors act in a competitive or coordinated fashion, having binding sites near or overlapping each other; thus, clustering of motifs can be exploited to detect higher confidence sites[7, 17, 18]. The ClusterBuster algorithm was designed for cluster detection of this sort, when provided with TF binding site matrices[7]. ClusterBuster defines a statistical model of a motif cluster, based on PWM models, and searches for sites in DNA that resemble the cluster model more than they resemble a model of 'background sequence' which is created by measuring the nucleotide abundances in the query sequences. Here ClusterBuster will be used to detect which TF binding sites reside in clusters of motifs.

The results show that a transcription factor motif found to be in a cluster of motifs is more likely to be a functional binding site (Figure 4). ClusterBuster's predictions perform approximately three to four times as well as simple weight matrix scan alone.

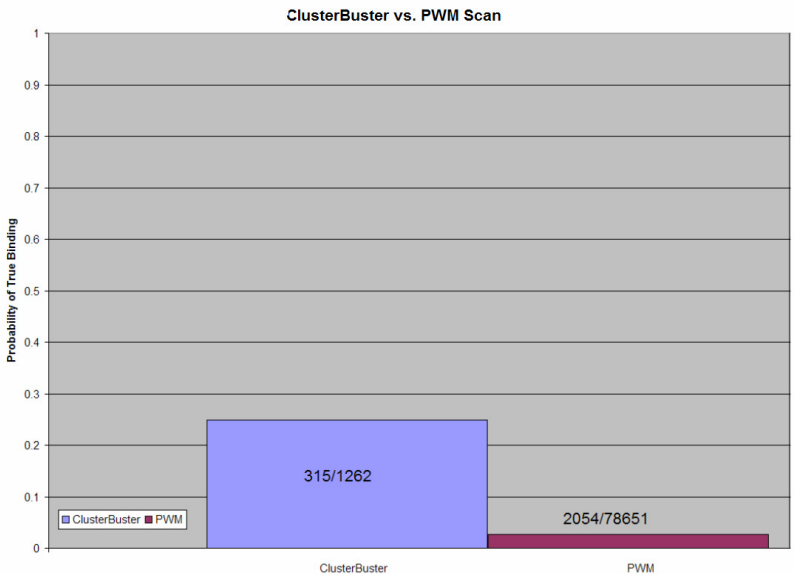


Figure 4 ClusterBuster: If a TF is present within a cluster of binding sites upstream of a gene the likelihood that the TF truly binds the gene is increased. ClusterBuster easily outperforms simple PSSM scans, meaning that the presence of a motif in a cluster of binding sites is more likely to be a true binding motif. Data is average over 104 TFs.

2.4 Expression Correlation as a Method to Predict Regulation

Analysis of MotifScanner outputs show that, not only is there a high rate of false positive hits, but there are also many missed binding sites. Expression analysis has the potential to discover targets missed by MotifScanner, as it finds regulatory relationships without prior cis-element searching. By definition, genes with similar expression are likely to be under similar regulatory pressures in the same way that genes regulated by the same TF are more likely to be co-expressed[19, 20]. Two thoughts are often pursued when deriving regulatory information from expression profile correlations. One is that transcription factors may regulate genes that have expression profiles similar to the TF[21]. The other is that groups of genes with similar expression may be regulated by the same TF(s).

Despite evidence of TF-target co-expression, it is clear that many transcription factors influence their targets, not by changes in their own expression, but by phosphorylation, nuclear exclusion, or some other mechanism. Binding site accessibility, modified by chromatin structures, is also a factor that alters gene expression. This kind of behavior can begin to be addressed by methods of gene expression clustering that do not rely on TF-target correlations. Considering the relevance of both types of TF-target prediction, both methods will be examined for their ability to predict regulation from microarrays.

TF-Target Correlations for Bayesian Allocation Measured by Profile Entropy Minimization

First, to elucidate how well TF-target co-expression indicates regulation, the correlation between each gene and every TF will be calculated using a recently developed strategy. True regulatory interactions can often be masked in large expression profiles due to noise in these data. Since genes are often regulated by different sets of TFs under different conditions, regulator-target relationships can be drowned out when expression profiles are sufficiently large. A new approach [22] addresses this problem by searching for the conditions under which a regulator's profile is maximally associated with a target's profile, essentially choosing the set of experiments where the TF most clearly and significantly controls the expression of a potential target. This has the advantage of removing the noise present under conditions where the target may be controlled by other factors, and allowing detection of correlations that would otherwise not be found if the entire expression profile was examined at once. In this analysis correlations with a p-value of 10^{-18} were chosen in order to extract the most significant regulatory relationships and reduce false predictions.

Target-Target Correlations-Dot Products of Expression Profile for SVM

For support vector machine classification the more simplistic vector dot product will be used to correlate gene expression, based on the expression vector of each gene across 1011 experiments[23]. This lends itself naturally to SVM classification since the dot product is a commonly used kernel function. Geometrically, the dot-product gives us information about the angle between two vectors. In this case, it can be interpreted as similarity between vectors, where the sign of the correlation tells us whether the correlation is positive or negative.

Co-regulated genes are expected to show similar expression patterns, and this will become evident through similarities shown in the kernel matrix. Given many known targets of a transcription factor as positive examples, the SVM can classify a new gene based on how closely its expression resembles that of the known examples.

2.5 Phylogenetic Profiles

Co-evolution of a transcription factor's targets may indicate regulation. A phylogenetic profile of a gene is

nothing more than the pattern of occurrence of its orthologs across a set of genomes. Genes with similar patterns have been shown to participate in the same physical complexes or have similar biochemical roles within the cell[24]. It has also been postulated that transcription factors and their targets co-evolve[25]. Therefore, it seems reasonable that a group of commonly regulated genes could share a similar pattern of inheritance. Phylogenetic profiles here were parsed from the COG database, which contains orthology information between *S.cerevisiae* and 65 other microbial genomes. Phylogenetic profiles were combined with other data for support vector machine classification as described below.

2.6 GO Annotation

Much like phylogenetic profiling, GO term annotation can be used to detect possible transcriptional targets. The targets of a transcription factor have often been shown to have similar function and a gene's GO annotation can be used to measure its functional similarity to known targets.

2.7 K-mer Distribution

PWMs may fail to detect binding sites if the binding site collection used to generate them was incomplete (in the case of experimental data) or if the motif discovery procedure was inaccurate (in the case of computationally generated matrices). In this case, 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. K-mer counts in promoters have been used before with SVMs to predict a gene's function[26]. Here, all promoters were decomposed into a vector of k-mers length 4, 5, and 6. Given a set of true positives and true negatives (discussed below), an SVM can classify a gene as a target or non-target based on its k-mer profile. K-mer profiles were generated using the program fasta2matrix[27].

2.6 Data Integration

Bayesian Allocation as a method of Data Integration

With the results of various sequence analysis methods in hand, the goal then becomes to integrate them into one predictive statistic. Integration of genomic data can be performed in several ways. Firstly, Bayesian probabilities can be exploited to determine high confidence site predictions. This can be done using an allocation method where data is compared to a set of known true sites, in this case a combination of ChIP-chip and literature data. In this way, a threshold can be set allowing each method to make predictions independently; or, given any combination of methods (e.g., degeneracy and conservation) predictions made from different data may overlap, thus bringing the true positives above the background. For example, most detected binding sites will not show repetition (i.e., only one site found in a promoter region) and so the probability that a single detected motif is a true binding site is low. However, single motifs which are also present in orthologous promoter regions have a higher likelihood of being real. In this analysis we take as predictions all interactions which fall into a category where at least 15% of interactions are true (known from positive set). Thus, the use of combined data is evident in a case where an interaction of degeneracy 1 will not meet the threshold (because the Degeneracy 1 category is not very predictive), but by combining conservation and degeneracy, the interaction may fall into Degeneracy-1 Conservation-4, which is a higher confidence category, passing the threshold. This methodology uses prior knowledge to choose divisions of the data which we believe contain better predictions. Note, however, that no negative examples are used as in SVM or naïve Bayes' classifiers.

It is clear that we cannot assume independence between each type of biological data. Thus, the relationship between the predicted and experimental data can be derived from Bayes' rule without assuming independence:

$$F(T \mid k_1) = \frac{F(k_1 \mid T)F(T)}{F(k_1 \mid T)F(T) + F(k_1 \mid \overline{T})F(\overline{T})}$$

where k_1 represents a particular category of data (e.g., 5 represents five repetitions of a motif for the Degeneracy method while 5 represents conservation in five genomes for the Conservation method), T indicates true binding, and T-bar means no binding. This representation is valid for the degeneracy data, conservation data, ClusterBuster data, and the expression data; and, in fact, all four of these methods (and potentially more) can be combined into one statistical measure by increasing the number of conditional parameters (i.e., k_1, k_2, k_3, …).

Thus, the probability that a TF binds a gene can be calculated taking into account several measures to enrich true binding sites. This type of analysis is very successful at reducing false positives and generating a small set of high confidence predictions. Since each individual method is somewhat successful at predicting regulation, the consensus of all measures together will isolate interactions that are highly likely to be true. The original results (on 104 TFs) using Bayesian allocation of the sort described did not produce very useful results. The result enriched for higher confidence sites as compared to PWM scan, but there were still many false hits and low sensitivity. This was thought to be because both the Degeneracy and Conservation Methods relied on PWMs, many of which appeared to be poor approximations of the true binding sites in that they were never found to recover even 5-10% of known true positives. To combat this, the analysis was performed again, after filtering out TFs whose

weight matrix could not recover at least 15% of known binding sites. The dataset overlaps identified for this smaller set of 62 TFs can be seen in figure 5. The data output from the Bayesian allocation analysis was tested using a split sample cross-validation procedure. The combination of methods produces predictions having greater precision than any one method, although sensitivity remains low compared to basic PWM scan (MotifScanner).

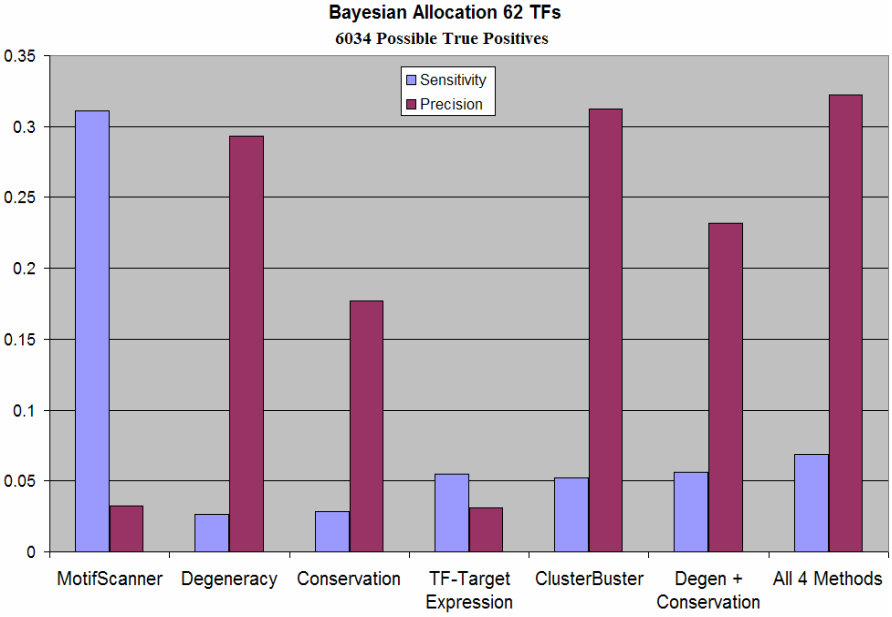


Figure 5 Bayesian Allocation on Degeneracy, Conservation, TF-Target Correlation, and ClusterBuster. Results are shown for the combination of Degeneracy and Conservation and the combination of all 4 methods together. The 4 Method combination produces the best results of any method alone. Raw MotifScanner data (PWM scan) is shown at far left for comparison. Motifscanner has a larger sensitivity but very low precison as it makes many false predictions.

Support Vector Machine as a Method of Data Integration

A support vector machine (SVM) is essentially a classification scheme for generating a linear classifier in some feature space defined by the data. The theory behind SVMs has been well explored, and many studies have been published applying SVMs to computational biology, primarily in classifying genes according to function[28-30]. In order to run an SVM based classification all data must be represented in a kernel matrix. Formally, this representation is discussed as an embedding of data into a feature space, F. Given some items of data, x_1 and x_2, their mapping can be shown as $\Phi(x_1)$ and $\Phi(x_2)$. Finally, some kernel function, $K(x_1, x_2) = <\Phi(x_1) , \Phi(x_2)>$ is used to specify the inner products of the data items in feature space. The result is a kernel matrix containing the inner products of every pair of data items. Practically speaking, no explicit mapping to a vector space is necessary since the classification is performed using only the inner products, which can be calculated directly from the data. The kernel matrix, though it must be generated using a valid kernel function, is really nothing more than a similarity matrix.

As with other classification techniques, a set of known positives and negatives must be supplied to the SVM. In this analysis, a classification will be done for each transcription factor independently (hence 104 separate classifiers). Positive examples for TF binding are taken from CHiP-Chip and other published experiments, and negatives are randomly chosen from those promoter regions that show no motifs for the particular TF under a very loose threshold of MotifScanner. For each method, every gene will have a list of attributes as its vector, and the scalar dot product between gene vectors will be the similarity measure used to create the unique kernel matrix for each data type (Degeneracy, Conservation, etc).

An example can be made using the data for Degeneracy. For motif frequency data let

$$\vec{D}^m = \left(D_1^m, D_2^m, D_3^m, ...D_n^m\right)$$

represent a set of TF motif frequencies for gene m, over n transcription factors (n=104). D_1^m would then be the number of times the motif for transcription factor 1 appears in the promoter of gene m.

Again, the dot product kernel function will be used to measure the similarity of gene vectors.

$dot = \vec{D}^m \bullet \vec{D}^k$ → denotes the dot product between vectors for genes m and k.

Once a kernel matrix is created for each method, they are easily combined simply by adding all matrices together. Other authors have explored several ways to combine data for SVM classification[26, 28]. It has been shown that combining data by generating individual kernels followed by simple matrix summation often gives the best performance (Figure 6). SVM classification was performed using the Gist software package[27], and results are reported after a leave-one-out cross-validation procedure.

ClusterBuster

	Motif 1	Motif 2	Motif 3
Gene1	1	1	0
Gene2	0	1	1
Gene3	0	0	1

Degeneracy

	Motif 1	Motif 2	Motif 3
Gene 1	2	1	2
Gene 2	0	0	1
Gene 3	1	0	1

Conservation

	Motif 1	Motif 2	Motif 3
Gene 1	4	8	0
Gene 2	5	0	0
Motif3	0	2	2

Phylogenetic Profile

	Genome1	Genome2	Genome3
Gene1	0.32	0.001	0.5
Gene2	-0.2	0.04	-0.001
Gene3	-0.6	0.4	-0.3

Expression

	Exp1	Exp2	Exp3
Gene1	0.32	0.001	0.5
Gene2	-0.2	0.04	-0.001
Gene3	-0.6	0.4	-0.3

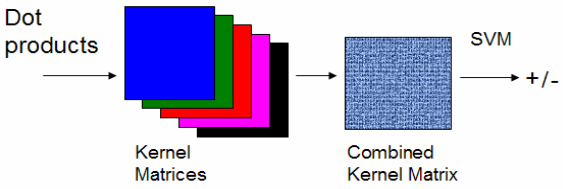


Figure 6 Separate Kernels are created for each method by taking the dot products of the gene vectors in each analysis. The various kernels are then added together to combine the various predictors. The composite kernel is then used for support vector classification.

Indeed Support Vector Classification performs better than simple Bayesian allocation and the combination of eight methods outperforms any one method by itself (Figure 7). SVM classification yields a sensitivity and accuracy not achieved by other methods.

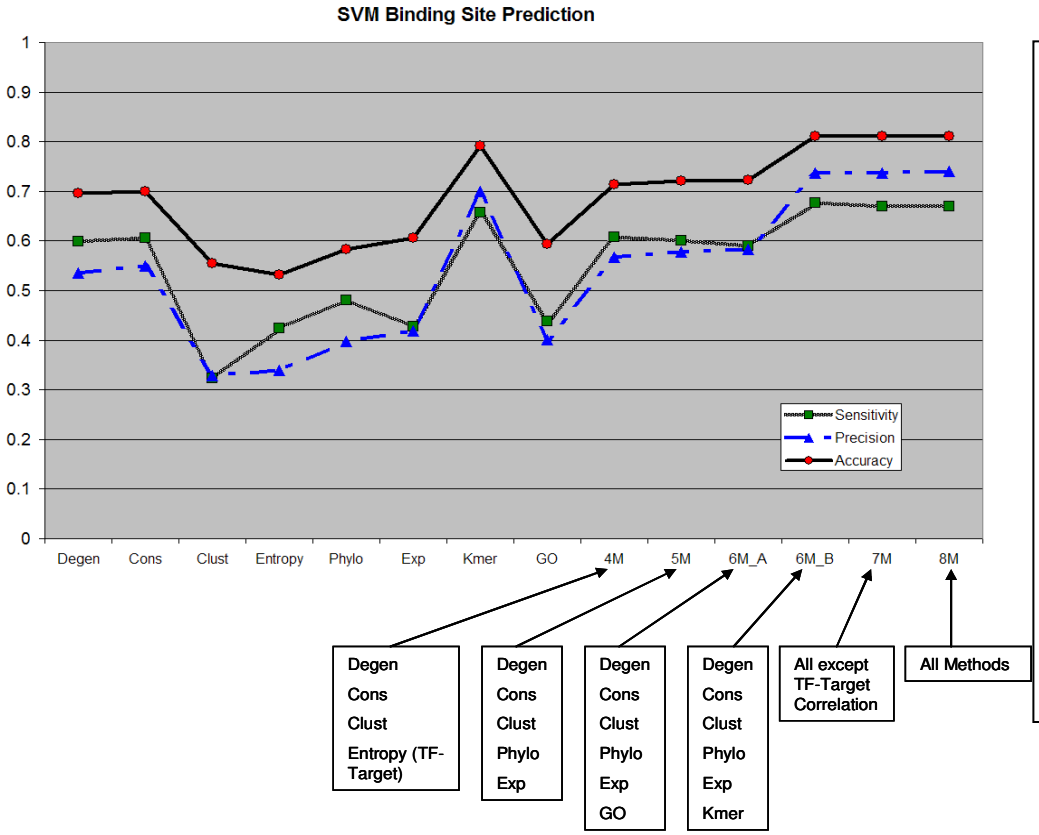


Figure 7 SVM Results SVM classification successfully integrates diverse biological data to make more accurate predictions than any single method alone.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

3 Discussions

A support vector machine offers several advantages when predicting binding sites. For instance, a gene vector for binding site degeneracy will contain the frequencies of binding site motifs for all TFs in the analysis, not just the one on which the classification is based. This inherently includes dependencies between transcription factors, something not taken into account with the Bayesian allocation procedure. If two or more TFs act together, then the positive examples will show predicted motifs for all of them rather than only one, reducing the ambiguity

when classifying an unknown gene. Furthermore, in the Bayesian example expression data had previously been modeled as a correlation between TF and target, meaning that if such a significant correlation was detected, the TF was predicted to bind the target. This only captures part of gene regulation since many TFs change their activity by post-translational modification or some other means. When creating a kernel matrix, each gene's expression profile forms a vector which can be compared to every other gene's profile using the scalar dot product. The resulting kernel is analogous to a Pearson correlation matrix where co-regulated genes are highly correlated. In this way, genes that have an expression profile similar to known targets will be classified as being bound by the TF. This is a much more intuitive measure since true targets are expected to have similar expression. Results for SVM classification using several types of genomic data are an improvement over other methods, providing increased sensitivity while reducing false positive identifications. Future work will include experimenting with different kernel functions to optimize classification, and comparing the SVM procedure to more conventional learning methods such as naïve Bayes, K-nearest-neighbor, and decision tree.

References

[1] C. Harbison, E. Fraenkel, R. Young and e. al., Transcriptional Regulatory Code of a Eukaryotic Genome, *Nature* 431 99-104,2004.

[2] I. T. Lee and e. al., Transcriptional Regulatory Networks in Saccharomyces cerevisiae, *Science* 298 799-804,2002.

[3] S. Aerts, G. Thijs, B. Coessens, M. Staes, Y. Moreau and B. De Moor, Toucan:Deciphering the Cis-Regulatory Logic of Coregulated Genes, *Nucleic Acids Research* 31 1753-1764,2003.

[4] V. Matys and e. al., TRANSFAC: Transciptional Regulation, from Patterns to Profiles, *Nucleic Acids Research* 31 374-378,2003.

[5] B. Ren, R. Young and et al., Genome-Wide Location and Function of DNA Binding Proteins, *Science* 290 2306-2309,2000.

[6] J. W. Fickett, Coordinate Positioning of MEF2 and Myogenin Binding Sites, *Gene* 172 19-32,1996.

[7] M. C. Frith, M. C. Li and Z. Weng, Cluster-Buster: Finding Dense Clusters of Motifs in DNA Sequences, *Nucleic Acids Research* 31 3666-3668,2003.

[8] M. Kellis and e. al., http://www.broad.mit.edu/annotation/fungi/comp_yeasts/, 2003.

[9] J. van Helden, Regulatory sequence analysis tools, *Nucleic Acids Research* 31 3593-3596,2003.

[10] B. Ewan and et al., An Overvie of Ensembl, *Genome Research* 14 925-928,2004.

[11] R. L. Tatusov and D. J. Lipman,dust,unpublished work.

[12] A. Smit and P. Green, Repeatmasker,

[13] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, W. Helmberg, D. L. Kenton, O. Khovayko, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, J. U. Pontius, K. D. Pruitt, G. D. Schuler, L. M. Schriml, E. Sequeira, S. T. Sherry, K. Sirotkin, G. Starchenko, T. O. Suzek, R. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko, Database resources of the National Center for Biotechnology Information, *Nucl. Acids Res.* 33 D39-45,2005.

[14] P. F. Cliften, M. Johnston and et al., Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting, *Science* 301 71-76,2003.

[15] M. Kellis, N. Patterson, M. Endrizzi, B. Birren and E. S. Lander, Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements, *Nature* 423 241-254,2003.

[16] P. F. Cliften and et al., http://www.genetics.wustl.edu/saccharomycesgenomes/, 2003.

[17] M. I. Amone and E. H. Davidson, The Hardwiring of Development: Organization and Function of Genomic Regulatory Systems, *Development* 124 1851-1864,1997.

[18] B. P. Berman and et al., Exploiting Transcription Factor Binding Site Clustering to Identify Cis-Regulatory Modules Involved in Pattern Formation in the Drosophila Genome, *Proceedings of National Academy of Science* 99 757-762,2002.

[19] H. Yu, N. Luscombe, J. Qian and M. Gerstein, Genomic Analysis fo Gene Expression Relationships in Transcriptional Regulatory Networks, *Trends in Genetics* 19 422-427,2003.

[20] D. Allocco, I. Kohane and A. Butte, Quantifying the Relationship Between Co-expression, Co-regulation, and Gene Function, *BMC Bioinformatics* 5 2004.

[21] Z. Zhu, Y. Pilpel and G. Church, Computational Identification of Transcription Factor Binding Sites via a Transcription-Factor-Centric-Clustering (TFCC) Algorithm, *Journal of Molecular Biology* 318 71-

81,2002.
[22] J. Mellor and C. DeLisi, Inferring the Logic of Context-Dependent Transcription Regulation, *Bioinformatics* In Press 2004.
[23] S. Bergman, J. Ihmels and N. Barkai, Iterative Signature Algorithm for the Analysis of Large-Scale Gene Expression Data, *Physical Review* 67 2003.
[24] J. Wu, S. Kasif and C. DeLisi, Identification of Functional Links Between Genes Using Phylogenetic Profiles, *Bioinformatics* 19 1-7,2003.
[25] A. Gasch, A. Moses, D. Chiang, H. Fraser, M. Berardini and M. Eisen, Conservation and Evolution of Cis-Regulatory Systems in Ascomycete Fungi, *PLOS Biology* 2 2202-2219,2004.
[26] P. Pavlidis and W. S. Noble, Gene Functional Classification from Heterogeneous Data, *RECOMB Conference Proceedings* 249-255,2001.
[27] P. Pavlidis, I. Wapinski and W. S. Noble, Support vector machine classification on the web, *Bioinformatics* 20 586-587,2004.
[28] W. S. Noble, Support Vector Machine Applications to Computational Biology, *Book Chapter* 2003.
[29] V. Vapnik, Statistical Learning Theory, *Text:The Nature of Statistical Learning Theory* 1998.
[30] V. Vapnik, The Nature of Statistical Learning Theory, *Text:The Nature of Statistical Learning Theory* 1999.