BIOINFORMATICS

Vol. 24 ISMB 2008, pages i165–i171
doi:10.1093/bioinformatics/btn154

Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences

Lucas D. Ward[1] and Harmen J. Bussemaker[1,2,*]

[1]Department of Biological Sciences, Columbia University, New York, NY 10027 and [2]Center for Computational Biology and Bioinformatics, Columbia University, New York, NY 10032, USA

ABSTRACT

Motivation: The identification of transcription factor (TF) binding sites and the regulatory circuitry that they define is currently an area of intense research. Data from whole-genome chromatin immunoprecipitation (ChIP–chip), whole-genome expression microarrays, and sequencing of multiple closely related genomes have all proven useful. By and large, existing methods treat the interpretation of functional data as a classification problem (between bound and unbound DNA), and the analysis of comparative data as a problem of local alignment (to recover phylogenetic footprints of presumably functional elements). Both of these approaches suffer from the inability to model and detect low-affinity binding sites, which have recently been shown to be abundant and functional.

Results: We have developed a method that discovers functional regulatory targets of TFs by predicting the total affinity of each promoter for those factors and then comparing that affinity across orthologous promoters in closely related species. At each promoter, we consider the minimum affinity among orthologs to be the fraction of the affinity that is functional. Because we calculate the affinity of the entire promoter, our method is independent of local alignment. By comparing with functional annotation information and gene expression data in *Saccharomyces cerevisiae*, we have validated that this biophysically motivated use of evolutionary conservation gives rise to dramatic improvement in prediction of regulatory connectivity and factor–factor interactions compared to the use of a single genome. We propose novel biological functions for several yeast TFs, including the factors Snt2 and Stb4, for which no function has been reported. Our affinity-based approach towards comparative genomics may allow a more quantitative analysis of the principles governing the evolution of non-coding DNA.

Availability: The MatrixREDUCE software package is available from http://www.bussemakerlab.org/software/MatrixREDUCE

Contact: Harmen.Bussemaker@columbia.edu

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

1 INTRODUCTION

Genome sequences encode not only the sequences of RNAs, but also the rates at which these are transcribed under various conditions. This *cis*-regulatory code is a consequence of the sequence specificity of transcription factors (TFs) and their interactions with other TFs, nucleosomes and other chromatin-associated proteins. The identification of *cis*-regulatory elements on a genomic scale is complicated by the fact that, although TF sequence specificity is generally well-characterized *in vitro*, functional elements in the genome are in fact much sparser than would be predicted from sequence alone (Gao *et al.*, 2004; Pilpel *et al.*, 2001). There are two types of constraint on the *in vivo* selection of functional targets by a given TF: those that prevent the TF from binding to DNA, and those that prevent a bound TF from driving transcription (Fig. 1). Functional and comparative genomics data are therefore needed in concert with knowledge of TF sequence specificity and DNA sequence to infer regulatory networks. The advantage of the functional genomics approach is that the contribution of sequence to the recruitment of a particular TF, and the association of that TF's binding with transcription, can be quantified. Comparative genomics, on the other hand, can lend evidence of the biological utility of TF binding through the application of evolutionary principles.

Most comparative genomics methods rely on local alignment of orthologous promoters or statistical measures of sequence overrepresentation (Cliften *et al.*, 2003; Kellis *et al.*, 2003; Li and Wong, 2005; Moses *et al.*, 2004; Siddharthan *et al.*, 2005; Sinha *et al.*, 2004), neither of which reflect the evolutionary constraints on regulatory sequence. Local alignment is not well-suited to detect lower affinity binding sites, which may be distant in sequence space yet functional; nor can it capture the rapid turnover of binding sites, which often occurs without conservation of position (Dermitzakis and Clark, 2002; Ludwig, 2002; Tautz, 2000; Wray, 2003). These limitations could be overcome by not directly comparing orthologous sequences, but rather comparing their predicted affinities for various TFs. Various biophysically motivated models of promoter-TF affinity have been developed (Bintu *et al.*, 2005; Djordjevic *et al.*, 2003; Liu and Clarke, 2002; Roider *et al.*, 2007; Ronen *et al.*, 2002; Stormo *et al.*, 1986) that allow such a comparison. In this article, we build upon the principles of conservation of promoter-TF affinity across a wide range of interaction strengths (Tanay, 2006) and conservation of the core transcriptional network (Pritsker *et al.*, 2004), and posit that the fraction of a promoter's affinity that is conserved in all species—that is, the minimum total affinity among orthologous promoters—can be used as a proxy for the fraction of the affinity that is functional.

We tested this idea using *Saccharomyces cerevisiae* and three closely related yeast species. We used the position-specific affinity matrix (PSAM) model (Foat *et al.*, 2006) to

[*]To whom correspondence should be addressed.

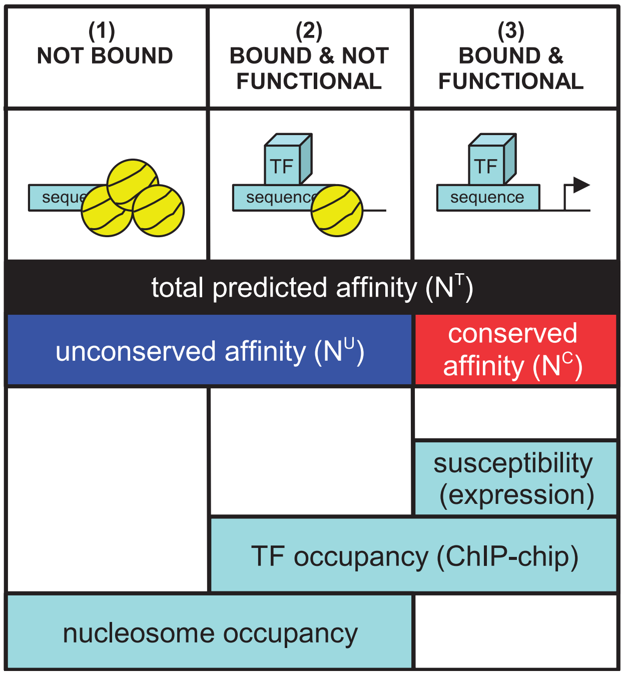

Fig. 1. Model for conservation of TF affinity. Promoter-TF affinity may be partitioned three ways: (**1**) affinity that does not lead to occupancy, (**2**) affinity that leads to occupancy but not function and (**3**) affinity that leads to occupancy and function. (1) and (2) might be expected to correlate with nucleosome occupancy, (2) and (3) with TF occupancy as measured by ChIP–chip and (3) with expression. Only (3) is expected to be conserved across orthologous promoters.

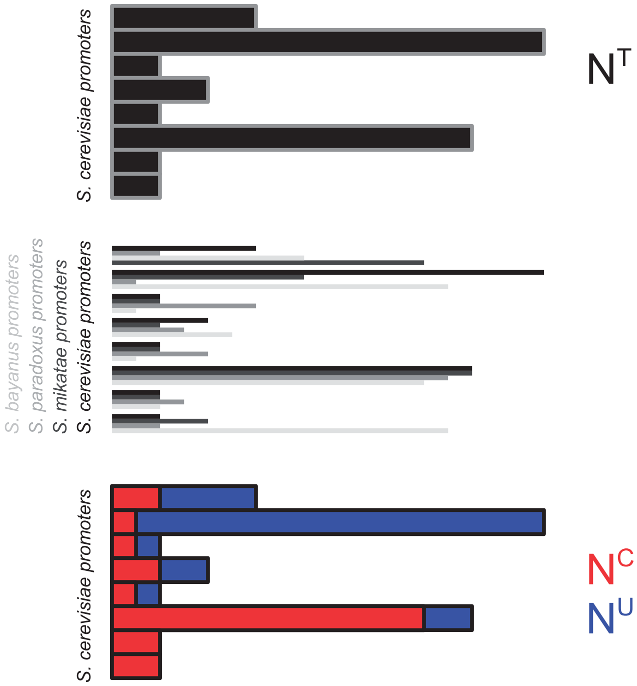

Fig. 2. Partitioning total affinity N^T at each promoter into a conserved fraction N^C and unconserved fraction N^U, where N^C is defined as the minimum predicted affinity across the four orthologs of each promoter, and $N^U = N^T - N^C$.

predict the affinity of each promoter in each species for a set of TFs with previously characterized sequence specificities (MacIsaac *et al.*, 2006). For each TF, this yields a value for the total affinity at every promoter in *S. cerevisiae* (N^T). The minimum of the four orthologous promoters' affinities defines the conserved promoter affinity (N^C) at each promoter, and the unconserved promoter affinity (N^U) is calculated by subtracting N^C from N^T (Fig. 2). We find that compared to the unconserved affinity N^U, the conserved affinity N^C tends to exhibit greater bias toward Gene Ontology (GO) categories, better explains TF-promoter susceptibilities inferred from expression data, and correlates more strongly with nucleosome depletion. For several TFs, we detect GO category enrichment using the conserved affinity N^C when none is observed using the total single-species affinity N^T and no function has been reported in the literature.

We also develop a measure of correlation between genome-wide N^C landscapes for pairs of TFs (affinity co-conservation). The interactions thus predicted are highly enriched for known physical or functional interactions between TFs. When the same approach is repeated using N^T (affinity co-occurrence), no such enrichment is detected.

Our method holds promise for predicting *in vivo* function when only *in vitro* TF binding data and an ensemble of closely related genome sequences are available. It is fundamentally different from other methods because it is free of the parameters which govern local alignment, thresholding between targets and non-targets, and any distinction between conserved and non-conserved instances of individual binding sites.

2 METHODS

2.1 Modeling TF affinities for orthologous promoter sequences

Genome sequences for *S. bayanus*, *S. cerevisiae*, *S. mikatae* and *S. paradoxus* were obtained from Kellis *et al.* (2003). Only genes for which the authors defined orthologs in all species were considered. For the correlation and GO analyses, we extracted sequences 500 bp upstream from each start codon and truncated them to exclude any upstream coding sequences. For the co-occurrence and co-conservation analyses, to negate length effects and spurious correlations, we extracted 500 bp upstream sequences without regard for upstream coding sequences; additionally, to avoid any overlap, we considered only the sequence upstream of the gene encoded on the Watson strand at divergent promoters with a length <1 kb.

We used the PSAM model for TF–DNA binding affinity (Foat *et al.*, 2006), consisting of a matrix of parameters w_{jb}. These weights represent the energetic consequences of mutations from the highest affinity sequence S_{ref} to nucleotide j at position b within the binding window, resulting in a mutated sequence S_{mut} with an increased free energy of binding:

$$w_{jb} = \frac{K_a(S_{\mathrm{mut}})}{K_a(S_{\mathrm{ref}})} = e^{\Delta\Delta G/RT} \qquad (1)$$

As described by Foat *et al.* (2007), assuming independence between positions, we can multiply the weights within a binding window to arrive at the relative affinity of the TF for any sequence S_{mut}. If we further assume a non-saturating physiological concentration of TF relative to DNA, the occupancy N of a sequence S_{mut} is directly proportional to this association constant K_a and the concentration of the factor [TF]. Finally, we sum the occupancies of each subsequence within a sliding window across a promoter to calculate the predicted occupancy N_g for a promoter sequence U_g of any length i:

$$N_g = [\mathrm{TF}]K_a(S_{\mathrm{ref}}) \sum_{i=1}^{L_g-L_w+1} \prod_{j=1}^{L_w} w_{jU_g(i+j-1)}. \tag{2}$$

We converted a library of position-specific scoring matrices (PSSMs) representing *S. cerevisiae* TF specificities from MacIsaac *et al.* (2006) to PSAMs using the transformation previously described by Foat *et al.* (2007) and Bussemaker *et al.* (2007b)

$$f_{jb} \propto (w_{jb})^{\lambda} p_b, \tag{3}$$

where λ is an evolutionary selection parameter (Berg and von Hippel, 1987; Stormo *et al.*, 1986), p_b is the background probability of base b, and f_{jb} is the frequency at position j and base b in the PSSM. Using a selection parameter $\lambda = 1$ and equal background frequencies p_b, we converted the $\log_2$-likelihood scores $s_{jb} = f_{jb}/p_b$ to weights $w_{jb} = 2^{s_{jb}}$. We then normalized each column of the PSAM so that the highest affinity base b at each position j was equal to one, and scaled the weights for the other bases accordingly.

We used the AffinityProfile utility from the *MatrixREDUCE* software package to calculate the affinity N_{gfs} for each promoter g for each TF f in each species s. We define the total promoter affinity N^T_{gf} as the affinity N_{gfs} for $s =$ *S. cerevisiae*, and the conserved promoter affinity (N^C_{gf}) of a promoter g for a TF f as the minimum of N_{gfs} among all four species s. The unconserved promoter affinity is given by

$$N^U_{gf} = N^C_{gf} - N^T_{gf}. \tag{4}$$

The genome-wide total promoter affinity conservation for each factor was calculated as

$$c_f = \frac{\sum_g N^C_{gf}}{\sum_g N^T_{gf}}. \tag{5}$$

2.2 Correlation with functional genomics data

We selected those TFs for which both a PSAM and a corresponding TF deletion expression microarray experiment from Hughes *et al.* (2000) were available and calculated the Pearson correlation between N^U for all promoters and the corresponding $\log_2$ expression ratios, and between N^C for all promoters and the corresponding $\log_2$ expression ratios.

Similarly, we considered TFs for which Gao *et al.* (2004) applied the *MA-Networker* algorithm, which predicts the regulatory susceptibility of each promoter to a given TF by combining whole-genome chromatin immunoprecipitation (ChIP–chip) data with a compendium of expression profiles. We calculated the Pearson correlation between both N^U and N^C for each TF and the corresponding promoter–TF coupling T-values from *MA-Networker*.

To investigate the relationship between conservation and nucleosome occupancy, we used the nucleosome ChIP–chip data from Bernstein *et al.* (2004). We calculated the Pearson correlation between either N^U or N^C for each TF on one hand, and the mean $\log_2$ ratios across nucleosome ChIP experiments on the other.

2.3 GO enrichment analysis

We performed GO (Ashburner *et al.*, 2000) analysis using a non-parametric variant of the *T-profiler* algorithm (Boorsma *et al.*, 2005) described by Scheer *et al.* (2006) and Bussemaker *et al.* (2007a). Briefly, using the N^T, N^C, and N^U for each TF, we performed a Mann–Whitney–Wilcoxon test between the affinities for genes within each GO category and affinities for all other genes. We performed a Bonferroni correction on the resulting P-values by multiplying them by the number of unique GO categories. We use the *YEAST* package from the BioConductor platform (Gentleman *et al.*, 2004) within the *R* statistical programming environment.

2.4 Affinity co-occurrence and co-conservation

We define the *affinity co-occurrence* between two TFs as the rank correlation between N^T profiles of those TFs beyond what is explained by the similarity between pairs of PSAMs. *Affinity co-conservation* is defined similarly, but using N^C instead of N^T. We first calculated the Spearman correlation coefficient ρ using each of the 7626 pairs of TFs. To control for the confounding effect of similarity between PSAMs, we then randomly permuted the sequence of every promoter in each genome, and performed the same analysis. Using 3874 random samples, we obtained a null distribution for ρ for each TF pair. This procedure was performed separately for co-occurrence and co-conservation. These distributions were found to be normal upon examination of a Q–Q plot, which then allowed us to calculate a mean and SD for the null distribution for each TF pair, and assign two P-values to each TF–TF pair which served as our co-occurrence and co-conservation metrics. We also assigned a false-discovery rate (FDR) threshold α to each pair as described by Benjamini and Hochberg (1995). We assigned ranks i to the co-affinity and co-occurrence P-values, and calculated the FDR threshold $\alpha = pn/i$, where n is the number of pairs (7626).

We compared our co-occurrence and co-conservation scores against four sources of validation: the subset of cofactor pairs reported by Banerjee and Zhang (2003) that had been experimentally validated; physical interactions deposited in the Saccharomyces Genome Database (SGD) (Cherry *et al.*, 1998), including interactions detected by mass spectrometry affinity capture, Western Blot affinity capture, complex reconstitution and yeast two-hybrid; synthetic lethal interactions from SGD; and synthetic rescue interactions from SGD. We used the Mann–Whitney–Wilcoxon P-value and the area under the ROC curve (ROC AUC) to assess the performance of the co-occurrence and co-conservation scores at predicting validated interactions of each type. We used the ROCR package developed by Sing *et al.* (2005) within the *R* statistical programming environment for calculating the ROC AUC.

3 RESULTS

3.1 Most predicted binding affinity for TFs is unconserved

For each TF, we first analyzed the overall conservation of affinity c_f, defined as the ratio of the genome-wide sum of conserved promoter affinities N^C and the genome-wide sum of total promoter affinities N^T. The overall conservation varies greatly between factors (Fig. 3), but in most cases, the majority of TF affinity is unconserved. This finding is consistent with previous observations that TF affinity is

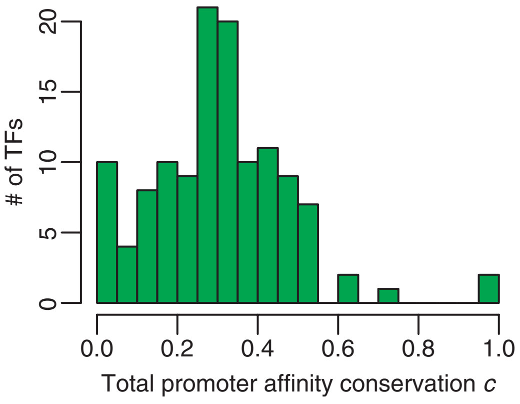

Fig. 3. Histogram displaying the genome-wide total promoter affinity conservation c_f for each of the studied factors.

not sufficient for TF binding, and that TF binding is not sufficient for function; in fact, previous work has shown that approximately 42% of TF binding is not functional (Gao *et al.*, 2004).

3.2 Conserved affinity correlates with *in vivo* regulatory susceptibility

To explore whether the conserved fraction of the affinity at a promoter N^C corresponds to the part of the affinity that is functional, and conversely whether the unconserved fraction N^U corresponds to non-functional affinity, we analyzed transcriptional response to TF deletion as measured using expression microarrays by Hughes *et al.* (2000). Using the Pearson correlation between both components of the affinity and the expression $\log_2$ ratios shows that, indeed, the N^C tends to predict the transcriptional response of a promoter better than the N^U (Fig. 4A).

Such microarray experiments are often difficult to interpret, because of the drastic physiological change induced and the lack of distinction between direct and indirect targets. The analysis by Gao *et al.* (2004) was one of a family of techniques to integrate binding and expression data to estimate the regulatory susceptibility of each *S. cerevisiae* promoter to each of a variety of TFs. The authors noted that this quantity did not correlate well with sequence-based predictions of affinity due to an abundance of non-functional binding. Again using the Pearson correlation, we found that these susceptibilities are explained by the conserved affinity N^C, in contrast to the unconserved affinity N^U (Fig. 4B).

3.3 Conserved affinity correlates with nucleosome depletion

There are a variety of reasons why a promoter with high affinity for a TF might not be regulated by that TF and consequently would not display conserved affinity across orthologous promoters. Specifically, (i) the TF may occupy the promoter, yet lack the appropriate cofactors at that promoter, or the preferred binding site within that promoter might be in the wrong position or orientation relative to the transcription start site to recruit or activate polymerase; or (ii) the TF might not bind as predicted by sequence because of competing occupancy by nucleosomes. We can distinguish between these two models using ChIP–chip data that probe genomic occupancy by nucleosomes and by the factors themselves. When we compare N^C and N^U for TFs with their corresponding promoter occupancies as measured by Harbison *et al.* (2004), we find modest support for the former mechanism; N^C and N^U tend to both correlate with binding of the TF, although N^C often correlates more strongly (data not shown). However, when we look at nucleosome occupancy data from Bernstein *et al.* (2004), we observe a strong and consistent relationship between nucleosome depletion and N^C, but not N^U (Fig. 5). Strikingly, the exceptions to this mechanism are Rap1, Sfp1, and Fhl1, all of which function to regulate ribosomal protein (RP) genes (Lieb *et al.*, 2001; Marion *et al.*, 2004; Yu and Morse, 1999).

3.4 Conserved affinity profiles allow predictions of novel TF functions

Any genome-wide measure of promoter–TF connectivity can be combined with prior classifications of genes such as those curated by the GO consortium (Ashburner *et al.*, 2000) to detect functional bias (Bussemaker *et al.*, 2007a). We tested the utility of calculating the conserved affinities N^C by analyzing their bias toward GO categories, and by comparing this bias to that which is detected using unconserved affinities N^U. We expected N^U to be distributed at random in the genome, and that any functional bias should be restricted to N^C. Indeed, we find that in general, the significance of the bias toward the most enriched GO category is much stronger when using N^C (Fig. 6). Again, the three exceptions are the factors Rap1, Sfp1, and Fhl1, which all regulate RP genes.

We also find that the number of GO categories within which significant bias is detected is almost always greater when using the conserved affinity N^C compared to the unconserved affinity N^U (Supplementary Table 1). In some cases, we are able to predict functions for TFs using N^C that are not ascertainable from N^T; these predictions tend to agree with those provided by the literature, when available (Supplementary Table 1). We are able to formulate

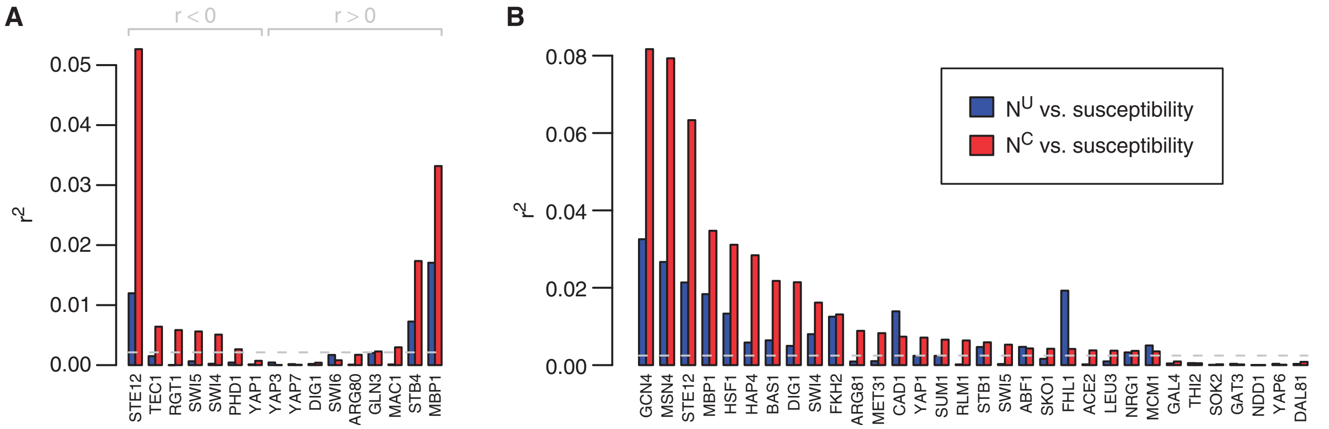

Fig. 4. Correlation of unconserved promoter affinity (N^U, blue) and conserved promoter affinity (N^C, red) with corresponding (**A**) deletion experiment expression data from Hughes *et al.* (2000) and (**B**) regulatory susceptibilities inferred from binding and expression across conditions by Gao *et al.* (2004). The gray dashed line indicates a significance threshold of $P=0.05$, Bonferroni-corrected for the number of correlation tests.

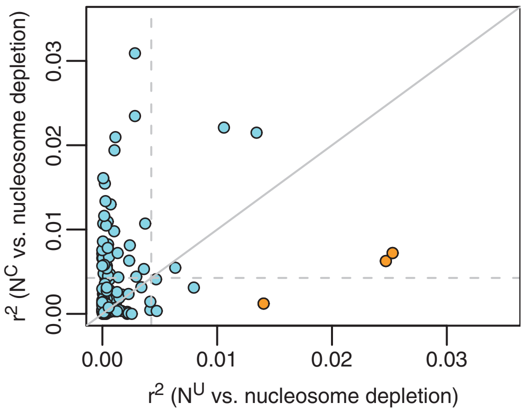

Fig. 5. Correlation of the conserved promoter affinity (N^C) and unconserved promoter affinity (N^U) for each TF with nucleosome depletion reported by Bernstein *et al.* (2004). Colored in orange are the proteins Rap1, Sfp1, and Fhl1, which regulate RP genes. The gray dashed lines indicate a significance threshold of $P = 0.05$, Bonferroni-corrected for the number of correlation tests; the gray solid line represents $y = x$.

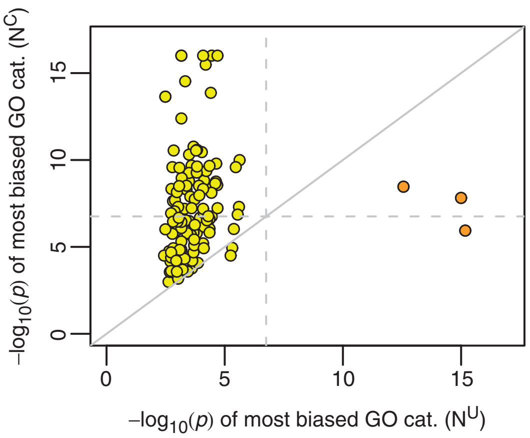

Fig. 6. Gene Ontology (GO) bias analysis using conserved promoter affinity (N^C) and unconserved promoter affinity (N^U). Uncorrected P-values are plotted; dashed lines represent a P-value threshold of 0.05, Bonferroni-corrected for the number of categories and factors. The gray solid line represents $y = x$. Colored in orange are the factors Rap1, Sfp1, and Fhl1, which regulate ribosomal protein (RP) genes.

several hypotheses about TF functions using N^C that have not been reported in the literature:

- Snt2, a DNA-binding protein with unknown function, displays bias in its N^C toward genes encoding amine transmembrane transporters ($P = 3.3 \times 10^{-3}$).
- Stb4, a DNA-binding protein with unknown function, displays bias in its N^C toward genes encoding transporters ($P = 1.5 \times 10^{-4}$).
- Yap6, a putative TF shown to be associated with salt tolerance (Fernandes *et al.*, 1997; Mendizabal *et al.*, 1998) displays bias in its N^C toward genes involved with carbohydrate metabolism ($P = 0.033$).
- The zinc-cluster protein encoded by YDR520C, which has been shown to be associated with caffeine sensitivity (Akache *et al.*, 2001), displays bias in its N^C toward genes involved with organic acid metabolism ($P = 2.5 \times 10^{-5}$).
- Yhp1 and Yox1 are both homeobox-containing transcriptional repressors that bind the cell-cycle regulator Mcm1 and release their repression on target genes at the M/G1 interval. Pramila *et al.* (2002) noted that the two factors' functions seem redundant as far as regulating known early cell-cycle target genes, and that they may interact with other cofactors to regulate distinct sets of genes. Indeed, we detect distinct conserved targets for the two factors: Yhp1 displays N^C biased toward genes involved with nucleotide and nucleotide triphosphate biosynthesis and metabolism ($P < 4.1 \times 10^{-3}$), and Yox1 appears targeted toward genes involved with the cell wall ($P = 4.8 \times 10^{-5}$).

3.5 Affinity for TFs that regulate RP genes displays a unique conservation pattern

In the course of our analysis, we made the surprising discovery that the factors Rap1, Sfp1, and Fhl1 do not follow the same pattern as the other studied factors: the unconserved affinity N^U for these factors appears to be functionally biased and associated with nucleosome depletion *more so* than the conserved affinity N^C. Each of these TFs is known to regulate RP genes, whose expression constitutes half of the RNA polymerase II transcription in rapidly growing yeast cells (Warner, 1999). Our GO analysis for these factors indicates that the strongest functional bias toward RPs in each case is found by considering N^T rather than either N^U or N^C. One explanation might be that the function of Rap1/Sfp1/Fhl1 affinity is context independent, leading even unconserved affinity to be functional, whereas for other TFs, a uniform background level of N^U is found throughout the genome (because it is non-functional outside of the proper genomic context, and therefore not selected against). The association of N^U with RP promoters, which one would expect to be exceptionally nucleosome depleted, explains the observed anomalous correlation between N^U for these factors and nucleosome depletion.

3.6 Affinity co-conservation provides evidence for TF–TF interactions

Co-occurrence and co-conservation of individual TF binding sites has been an extensively studied line of evidence for predicting cofactor pairs (Chiang *et al.*, 2003; Pilpel *et al.*, 2001; Sudarsanam *et al.*, 2002). Pairwise comparison of genome-wide N^T and N^C distributions constitutes a natural extension to these methods, which we term *affinity co-occurrence* and *affinity co-conservation*. Similarity in genome-wide N^T and N^C profiles between two TFs is governed, to an extent, by the similarity between their sequence specificities (PSAMs); we controlled for this effect by developing a null model for each TF–TF pair based on scrambled genomes. At a FDR threshold of $\alpha < 0.05$, we discovered 29 pairs using co-occurrence and 1530 by co-conservation (reported in full in Supplementary Table 2). We first compared these results to the subset of 11 cofactors predicted by Banerjee and Zhang (2003) that were experimentally validated. None of these validated pairs were among the 29 co-occurring pairs, while 8 of the 11 validated pairs were among the 1530 co-conserved pairs (hypergeometric $P = 1.9 \times 10^{-5}$). At this FDR-level stringency it is difficult to compare the two methods, so we used the ROC AUC (Fig. 7) and Wilcoxon-Mann-Whitney tests to compare their performance at predicting the 11 validated pairs. Strikingly, affinity co-occurrence fails to predict these cofactors by either measure, while affinity co-conservation predicts them with significant strength.

We then compared our predictions to other sets of TF interactions discovered via high-throughput experiments: physical interactions, synthetic lethal interactions, and synthetic rescue interactions (Table 1).

While physical interaction or synthetic lethality between two TFs does not necessarily imply that they act together as cofactors or even target the same genes, we nevertheless find that these types of evidence are associated with significant signals of affinity co-conservation. Synthetic rescue interactions are not associated with co-conservation, which is to be expected because such interactions indicate a complementary rather than synergistic relationship. Again, it is notable that in none of these cases does co-occurrence of affinity in a single genome provide evidence that agrees with experimentally validated pairs.

The pairs of interactions that we detect by co-conservation (Supplementary Table 2) are biologically plausible. The master cell-cycle regulators Fkh1 and Fkh2 have many interaction partners; the cell-cycle regulators Mcm1 and Yox1 are connected; the RP regulators Fhl1, Rap1, and Sfp1 are connected; heme-activator proteins Hap1, Hap2, Hap3, and Hap4 are connected; sulfur amino acid biosynthesis regulators Met4 and Met32 are connected; and glycolysis regulators Gcr1 and Gcr2 are connected. Interestingly, the serum response factor like protein Rlm1 has many interaction partners, suggesting that it has a genome-wide regulatory function that is shared by many factors with diverse functions.

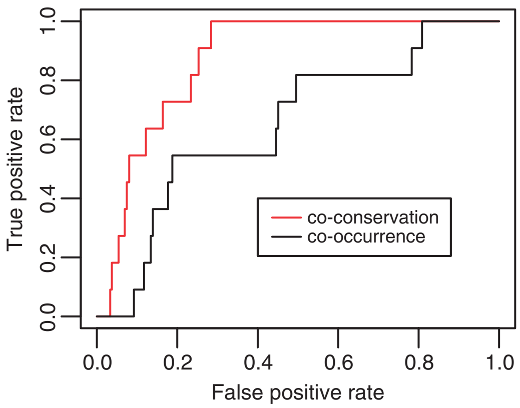

Fig. 7. ROC curves comparing the performance of the co-occurrence (black line) and co-conservation (red line) methods at predicting the cofactor pairs reported by Banerjee and Zhang (2003).

Table 1. Comparison of performance of the co-occurrence and co-conservation methods

Validation source	Co-occurrence ROC AUC	 Wilcoxon *P*	Co-conservation ROC AUC	 Wilcoxon *P*
Known cofactors[a]	0.65	0.08	0.87	1.9×10^{-5}
Physical interactions	0.52	0.64	0.59	7.1×10^{-3}
Synthetic lethal	0.66	0.09	0.73	1.7×10^{-2}
Synthetic rescue	0.53	0.66	0.47	0.69

[a] Banerjee and Zhang (2003).

CONCLUSION

We have exploited the pattern of TF affinity conservation across orthologous promoters to infer the fraction of TF affinity at each promoter that is functional. This method should be broadly applicable to situations where the *in vitro* binding specificity of a TF is known, but its *in vivo* function has not been demonstrated experimentally. It may also be especially useful in the analysis of the genomes of higher eukaryotes, in which non-functional binding sites outnumber functional ones to an even higher degree than in yeast. In contrast to phylogenetic footprinting methods, information from the full range of potential binding affinities is incorporated, allowing for more sensitive detection of potential TF–promoter and TF–TF interactions.

ACKNOWLEDGEMENTS

The author would like to thank Richard Mann, Gabor Halasz, and Ronald Tepper for critical readings of the manuscript, and members of the Bussemaker laboratory for helpful discussions.

Funding: This work was supported by National Institutes of Health (NIH) grants R24GM074105, T32GM008798, R01HG003008, and U54CA121852.

Conflict of Interest: none declared.

REFERENCES

Akache,B. *et al.* (2001) Phenotypic analysis of genes encoding yeast zinc cluster proteins. *Nucleic Acids Res.*, **29**, 2181–2190.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.*, **25**, 25–29.

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

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.

Berg,O.G. and von Hippel,P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, **193**, 723–750.

Bernstein,B.E. *et al.* (2004) Global nucleosome occupancy in yeast. *Genome Biol.*, **5**, R62.

Bintu,L. *et al.* (2005) Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*, **15**, 116–124.

Boorsma,A. *et al.* (2005) T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.*, **33**, W592–W595.

Bussemaker,H.J. *et al.* (2007a) Dissecting complex transcriptional responses using pathway-level scores based on prior information. *BMC Bioinformatics*, **8** (Suppl 6), S6.

Bussemaker,H.J. *et al.* (2007b) Predictive modeling of genome-wide mRNA expression: from modules to molecules. *Annu. Rev. Biophys. Biomol. Struct.*, **36**, 329–347.

Cherry,J.M. *et al.* (1998) SGD: Saccharomyces Genome Database. *Nucleic Acids Res.*, **26**, 73–79.

Chiang,D.Y. *et al.* (2003) Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol.*, **4**, R43.

Cliften,P. *et al*. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, **301**, 71–76.

Dermitzakis,E.T. and Clark,A.G. (2002) Evolution of transcription factor binding sites in mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.*, **19**, 1114–1121.

Djordjevic,M. *et al*. (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res.*, **13**, 2381–2390.

Fernandes,L. *et al*. (1997) Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Mol. Cell Biol.*, **17**, 6982–6993.

Foat,B.C. *et al*. (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics*, **22**, e141–e149.

Foat,B.C. *et al*. (2007) TransfactomeDB: a resource for exploring the nucleotide sequence specificity and condition-specific regulatory activity of trans-acting factors. *Nucleic Acids Res.*

Gao,F. *et al*. (2004) Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, **5**, 31.

Gentleman,R.C. *et al*. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

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

Hughes,T.R. *et al*. (2000) Functional discovery via a compendium of expression profiles. *Cell*, **102**, 109–126.

Kellis,M. *et al*. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Li,X. and Wong,W.H. (2005). Sampling motifs on phylogenetic trees. *Proc. Natl Acad. Sci. USA*, **102**, 9481–9486.

Lieb,J.D. *et al*. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat. Genet.*, **28**, 327–334.

Liu,X. and Clarke,N.D. (2002) Rationalization of gene regulation by a eukaryotic transcription factor: calculation of regulatory region occupancy from predicted binding affinities. *J. Mol. Biol.*, **323**, 1–8.

Ludwig,M.Z. (2002) Functional evolution of noncoding DNA. *Curr. Opin. Genet. Dev.*, **12**, 634–639.

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

Marion,R.M. *et al*. (2004) Sfp1 is a stress- and nutrient-sensitive regulator of ribosomal protein gene expression. *Proc. Natl Acad. Sci. USA*, **101**, 14315–14322.

Mendizabal,I. *et al*. (1998) Yeast putative transcription factors involved in salt tolerance. *FEBS Lett.*, **425**, 323–328.

Moses,A.M. *et al*. (2004) MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. *Genome Biol.*, **5**, R98.

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

Pramila,T. *et al*. (2002) Conserved homeodomain proteins interact with MADS box protein Mcm1 to restrict ECB-dependent transcription to the M/G1 phase of the cell cycle. *Genes Dev.*, **16**, 3034–3045.

Pritsker,M. *et al*. (2004) Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.*, **14**, 99–108.

Roider,H.G. *et al*. (2007) Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, **23**, 134–141.

Ronen,M. *et al*. (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc. Natl Acad. Sci. USA*, **99**, 10555–10560.

Scheer,M. *et al*. (2006) JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. *Nucleic Acids Res.*, **34**, W510–W515.

Siddharthan,R. *et al*. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.*, **1**, e67.

Sing,T. *et al*. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

Sinha,S. *et al*. (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila. *BMC Bioinformatics*, **5**, 129.

Stormo,G.D. *et al*. (1986) Quantitative analysis of the relationship between nucleotide sequence and functional activity. *Nucleic Acids Res.*, **14**, 6661–6679.

Sudarsanam,P. *et al*. (2002) Genome-wide co-occurrence of promoter elements reveals a cis-regulatory cassette of rRNA transcription motifs in Saccharomyces cerevisiae. *Genome Res.*, **12**, 1723–1731.

Tanay,A. (2006) Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.*, **16**, 962–972.

Tautz,D. (2000) Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.*, **10**, 575–579.

Warner,J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437–440.

Wray,G.A. (2003) Transcriptional regulation and the evolution of development. *Int. J. Dev. Biol.*, **47**, 675–684.

Yu,L. and Morse,R.H. (1999) Chromatin opening and transactivator potentiation by Rap1 in Saccharomyces cerevisiae. *Mol. Cell Biol.*, **19**, 5279–5288.