2005

Identification of functional transcription factor binding sites using closely related Saccharomyces species

Scott W. Doniger
Washington University School of Medicine in St. Louis

Juyoung Huh
Washington University School of Medicine in St. Louis

Justin C. Fay
Washington University School of Medicine in St. Louis

Follow this and additional works at: http://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation
http://digitalcommons.wustl.edu/open_access_pubs/2086

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Identification of functional transcription factor binding sites using closely related *Saccharomyces* species

Scott W. Doniger, Juyoung Huh and Justin C. Fay

*Genome Res.* 2005 15: 701-709
Access the most recent version at doi:10.1101/gr.3578205

---

**Supplemental Material**
http://genome.cshlp.org/content/suppl/2005/04/18/gr.3578205.DC1.html

**References**
This article cites 48 articles, 32 of which can be accessed free at:
http://genome.cshlp.org/content/15/5/701.full.html#ref-list-1

**Open Access**
Freely available online through the *Genome Research* Open Access option.

**Creative Commons License**
This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email Alerting Service**
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.

---

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Cold Spring Harbor Laboratory Press
Identification of functional transcription factor binding sites using closely related Saccharomyces species

Scott W. Doniger,1 Juyoung Huh,2 and Justin C. Fay1,2,3

1Computational Biology Program and 2Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA

Comparative genomics provides a rapid means of identifying functional DNA elements by their sequence conservation between species. Transcription factor binding sites (TFBSs) may constitute a significant fraction of these conserved sequences, but the annotation of specific TFBSs is complicated by the fact that these short, degenerate sequences may frequently be conserved by chance rather than functional constraint. To identify intergenic sequences that function as TFBSs, we calculated the probability of binding site conservation between Saccharomyces cerevisiae and its two closest relatives under a neutral model of evolution. We found that this probability is <5% for 134 of 163 transcription factor binding motifs, implying that we can reliably annotate binding sites for the majority of these transcription factors by conservation alone. Although our annotation relies on a number of assumptions, mutations in five of five conserved Ume6 binding sites and three of four conserved Ndt80 binding sites show Ume6- and Ndt80-dependent effects on gene expression. We also found that three of five unconserved Ndt80 binding sites show Ndt80-dependent effects on gene expression. Together these data imply that although sequence conservation can be reliably used to predict functional TFBSs, unconserved sequences might also make a significant contribution to a species’ biology.

[Supplemental material is available online at www.genome.org.]
that conservation among three closely related species is sufficient to predict functional TFBSs, making it possible to annotate the genome for functionally constrained binding sites for the majority of known transcription factors. Surprisingly, our annotation suggests that the TFBSs account for less than half of the functional constraint in noncoding sequences.

Results

Functional constraint in intergenic sequences

The identification of TFBSs by their sequence conservation between species requires that nonfunctional or unconstrained TFBSs are rarely conserved by chance. The probability that a nonfunctional binding site is conserved depends on the neutral substitution rate. We estimated the neutral substitution rate along the lineages leading to \textit{S. cerevisiae}, \textit{S. paradoxus}, and \textit{S. mikatae} from the median synonymous substitution rate of 2098 coding sequence alignments of these species (Table 1). The total rate across the phylogeny is 0.83 substitutions per site and is a conservative estimate of the neutral rate, since synonymous sites have been shown to be under weak selective constraint (Akashi 2001). At this distance, the probability that a 10-base pair sequence is identical across the three species is 0.002 (Kimura 1980). This implies that TFBSs may rarely be conserved between these three species by chance.

The fraction of functionally constrained intergenic sequences can be estimated from the ratio of the intergenic to synonymous substitution rate (Table 1) (Wong and Nielsen 2004). From 4188 alignments, the median intergenic substitution rate across the three yeast species is 0.57, which implies that 43% of intergenic sequences are functionally constrained. The extent to which conserved TFBSs can account for this constraint is discussed below.

Identification of functionally constrained Ndt80 and Ume6 binding sites using a neutral model of molecular evolution

The neutral substitution rate implies that functional TFBSs can be identified by sequence conservation alone, regardless of the conservation of the flanking sequences. We tested this hypothesis using two well characterized transcription factors, Ndt80 and Ume6. Both proteins regulate the expression of meiosis-specific genes in \textit{S. cerevisiae}. Ume6 is known to repress genes during vegetative growth and activate genes during early meiosis (Boudish and Mitchell 1993; Steber and Esposito 1995). Ume6 affects gene expression by binding to the consensus sequence TSGGC GGCTAW (Williams et al. 2002). Ndt80 activates genes expressed in the middle stages of meiosis (Chu and Herskowitz 1998) by binding to the consensus sequence YGNCACAAAW (Pierce et al. 2003). We used a position weight matrix (PWM) representation of these TFBSs creating a probabilistic description of the nucleotide frequencies at each position of the motif. Our goal was to determine which sites matching these sequences function in Ndt80 or Ume6 transcriptional regulation. If functional Ndt80 and Ume6 binding sites are also functional in other species, then they are likely to be under purifying selection and can be identified in \textit{S. cerevisiae} by their conservation at orthologous positions in other \textit{Saccharomyces} species.

For each site identified as a match to the Ndt80 or Ume6 PWM (Hertz and Stormo 1999), we counted the number of differences observed between \textit{S. cerevisiae}, \textit{S. paradoxus}, and \textit{S. mikatae}. The neutral model predicts that there will be an average of 4.8 and 5.3 differences across the three species for Ndt80 and Ume6, respectively, and that less than 1.5% of the Ndt80 and 0.8% of the Ume6 sites should have 0 or 1 difference between the three species. For both transcription factors, we found a substantial overrepresentation of sites with 0 or 1 difference, 63% of Ndt80 sites and 97% of Ume6 sites, suggesting that only a small number of the Ndt80 sites or Ume6 sites are consistent with a neutral model (Fig. 1). When we restricted the analysis to only those sites occurring in promoters of genes that are likely targets of these transcription factors (Chu et al. 1998; Williams et al. 2002), there was an even further enrichment for constrained sites (Fig. 1). These results imply that we can readily distinguish constrained and unconstrained sites using the evolutionary data from these three species.

![Figure 1](image-url)
To find the conserved TFBSs in the genome, we identified all sites that had a significant match to the PWM at orthologous positions in the three species. The advantage of this conservation test is that it does not rely on sequence conservation, which is not necessarily equal to TFBS conservation due to the degenerate nature of TFBSs (Moses et al. 2003). For example, we found that six of 20 Ndt80 sites and eight of 21 Ume6 sites with a single difference were not actually conserved sites, in that the sequence in either the \textit{S. paradoxus} or \textit{S. mikatae} genome was no longer a significant match to the PWM. Additionally, we identified a match to the Ndt80 site in the \textit{GAS4} promoter of \textit{S. cerevisiae} that has three differences across the three species, but the sequence in each species is a match to the Ndt80 PWM. In total, we identified 59 conserved Ndt80 TFBSs and 63 conserved Ume6 TFBSs. The list of genes with conserved Ndt80 and Ume6 TFBSs can be found in the Supplemental material (Tables S1 and S2).

The fraction of conserved Ndt80 and Ume6 TFBSs that are functional depends on how often a TFBS is conserved in the absence of selective constraint. We modeled the probability of TFBS conservation as the likelihood that a TFBS observed in \textit{S. cerevisiae} remains a match to the PWM given the neutral rate of evolution of 0.83 substitutions per site. Because there can be a number of sequences that match a PWM, we enumerated all possible evolutionary descendents of the observed sequence and tested each sequence for a significant match to the PWM. If it matched, we calculated the probability of it occurring given the neutral substitution rate. The total probability that a TFBS is conserved is the cumulative probability of all evolutionary paths, conditioned on the initial sequence and the evolutionary time between them.

Mutation of conserved Ndt80 and Ume6 binding sites produces Ndt80- and Ume6-dependent effects on gene expression

The conservation of Ndt80 and Ume6 binding sites suggests that they are true TFBSs and should therefore affect the expression levels of their adjacent genes. To test our predictions, we randomly selected four promoters containing conserved Ndt80 sites and four promoters containing conserved Ume6 sites (Fig. 3). We cloned each intergenic sequence, averaging 625 base pairs, and made mutations in the conserved TFBSs, changing the Ndt80 sites from YGNCACAAAW to YGNCTTCAAAW, and the Ume6 sites from TSGGCGGCTAW to TSGATTGCTAW. The expression of the mutant and wild-type promoters were compared using a β-gal reporter construct expressed from a yeast episomal plasmid. One of the promoters selected, the region upstream of \textit{ADY2}, contains two conserved Ume6 TFBSs, and mutations were made in both and tested independently.

Eight of nine mutants produced a significant change in expression levels during meiosis compared to the wild-type promoters from which they were derived (Fig. 4). Mutations in the

To find the conserved TFBSs in the genome, we identified all sites that had a significant match to the PWM at orthologous positions in the three species. The advantage of this conservation test is that it does not rely on sequence conservation, which is not necessarily equal to TFBS conservation due to the degenerate nature of TFBSs (Moses et al. 2003). For example, we found that six of 20 Ndt80 sites and eight of 21 Ume6 sites with a single difference were not actually conserved sites, in that the sequence in either the \textit{S. paradoxus} or \textit{S. mikatae} genome was no longer a significant match to the PWM. Additionally, we identified a match to the Ndt80 site in the \textit{GAS4} promoter of \textit{S. cerevisiae} that has three differences across the three species, but the sequence in each species is a match to the Ndt80 PWM. In total, we identified 59 conserved Ndt80 TFBSs and 63 conserved Ume6 TFBSs. The list of genes with conserved Ndt80 and Ume6 TFBSs can be found in the Supplemental material (Tables S1 and S2).

The fraction of conserved Ndt80 and Ume6 TFBSs that are functional depends on how often a TFBS is conserved in the absence of selective constraint. We modeled the probability of TFBS conservation as the likelihood that a TFBS observed in \textit{S. cerevisiae} remains a match to the PWM given the neutral rate of evolution of 0.83 substitutions per site. Because there can be a number of sequences that match a PWM, we enumerated all possible evolutionary descendents of the observed sequence and tested each sequence for a significant match to the PWM. If it matched, we calculated the probability of it occurring given the neutral substitution rate. The total probability that a TFBS is conserved is the cumulative probability of all evolutionary paths, conditioned on the initial sequence and the evolutionary time between them.

Mutation of conserved Ndt80 and Ume6 binding sites produces Ndt80- and Ume6-dependent effects on gene expression

The conservation of Ndt80 and Ume6 binding sites suggests that they are true TFBSs and should therefore affect the expression levels of their adjacent genes. To test our predictions, we randomly selected four promoters containing conserved Ndt80 sites and four promoters containing conserved Ume6 sites (Fig. 3). We cloned each intergenic sequence, averaging 625 base pairs, and made mutations in the conserved TFBSs, changing the Ndt80 sites from YGNCACAAAW to YGNCTTCAAAW, and the Ume6 sites from TSGGCGGCTAW to TSGATTGCTAW. The expression of the mutant and wild-type promoters were compared using a β-gal reporter construct expressed from a yeast episomal plasmid. One of the promoters selected, the region upstream of \textit{ADY2}, contains two conserved Ume6 TFBSs, and mutations were made in both and tested independently.

Eight of nine mutants produced a significant change in expression levels during meiosis compared to the wild-type promoters from which they were derived (Fig. 4). Mutations in the
four Ndt80 sites in sporulation media resulted in a 50-fold decrease in gene expression on average, consistent with previously characterized Ndt80 binding sites (Ozsarac et al. 1997; Chu et al. 1998). Mutations in two of the Ume6 sites resulted in an average eightfold increase in expression from the mutant promoter, whereas mutations in the three others produced an average 100-fold decrease in gene expression levels. These data are consistent with previous studies that showed that Ume6 can act as both a repressor and activator of gene expression (Bowdish and Mitchell 1993). Interestingly, the two Ume6 binding sites within the ADY2 promoter have an opposite effect on expression levels, despite having identical sequences.

If the conserved Ndt80 and Ume6 binding sites are functional Ndt80 and Ume6 binding sites, the effects of mutations within these sites should be Ndt80- and Ume6-dependent. We measured expression levels in either an ndt80Δ or ume6Δ strain where appropriate (Fig. 4). Seven of the nine mutant promoters show no significant difference compared to wild type in the absence of the transcription factor. Surprisingly, the mutant Ndt80 site in YFL040W, the one promoter that showed no significant difference in the absence of Ndt80, showed a significant increase in gene expression in the ndt80Δ strain in complete media (Supplemental Fig. 2). Further examination of this promoter revealed that there is a conserved Sum1 binding site overlapping the Ndt80 site. Ndt80 and Sum1 compete for binding to similar sites in the genome, and Sum1 acts to repress mitotic genes during vegetative growth (Pierce et al. 2003). The difference between the mutant and wild-type promoters may be caused by disruption of Sum1 binding, but only in the absence of Ndt80. The two additional sites which showed a significant difference in the deletion strains are the Ume6 binding sites in the ADY2 promoter. These mutant sites, which showed opposing effects in the S288C background, showed similar small effects in the absence of Ume6.

Mutations in Ndt80 binding sites that are not conserved affect gene expression

One assumption underlying the comparative genomic approach is that functional TFBSs will be conserved between closely related species (Cliften et al. 2003; Kellis et al. 2003). To test the importance of conservation in identifying functional regulatory sites, we tested five Ndt80 binding sites that are not conserved between the three species for Ndt80-specific regulatory effects (Supplemental Fig. 1).

Of five unconserved sites, two (URA4 and YMR111C) showed an Ndt80-specific effect in sporulation medium, suggesting that these sites are functional Ndt80 binding sites in S. cerevisiae, despite their absence in the other species (Fig. 5). One additional site, YBR255C-A showed a significant effect on gene expression in complete media, but showed no significant effect in sporulation media (Supplemental Fig. 3). The remaining two sites did not alter transcription when mutated in the S288C background.

Genome annotation of transcription factor binding sites

The low probability of TFBS conservation under a neutral model combined with the experimental validation of multiple Ndt80 and Ume6 binding sites suggests that the same method may be applied to all other known S. cerevisiae transcription factor binding motifs. We compiled a list of 163 unique motifs from three

---

**Table 2. Binding site conservation for five known transcription factors**

<table>
<thead>
<tr>
<th>Name</th>
<th>Consensus</th>
<th>S. cer. sitesa</th>
<th>S. par. sitesa</th>
<th>S. mit. sitesa</th>
<th>Conserved sitesb</th>
<th>Percent conservedc</th>
<th>P(cons)d</th>
<th>False discovery rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABF1</td>
<td>RTCRYNNNNNACG</td>
<td>671</td>
<td>670</td>
<td>666</td>
<td>470</td>
<td>70.0</td>
<td>0.017</td>
<td>0.024</td>
</tr>
<tr>
<td>GAL4</td>
<td>CGGNNNNNNNNNNNCGG</td>
<td>144</td>
<td>152</td>
<td>134</td>
<td>21</td>
<td>14.6</td>
<td>0.022</td>
<td>0.133</td>
</tr>
<tr>
<td>MCM1</td>
<td>CCNNNWWRGG</td>
<td>963</td>
<td>1019</td>
<td>911</td>
<td>194</td>
<td>20.1</td>
<td>0.023</td>
<td>0.112</td>
</tr>
<tr>
<td>NDT80</td>
<td>YGNCAAAAW</td>
<td>141</td>
<td>133</td>
<td>141</td>
<td>59</td>
<td>41.8</td>
<td>0.006</td>
<td>0.015</td>
</tr>
<tr>
<td>UME6</td>
<td>TSGCGCCTAW</td>
<td>82</td>
<td>86</td>
<td>86</td>
<td>63</td>
<td>76.8</td>
<td>0.002</td>
<td>0.003</td>
</tr>
</tbody>
</table>

aThe number of matches to the PWM in each genome.
bThe number of conserved matches to the PWM.
cThe percentage of matches to the PWM that are conserved.
dThe probability of motif conservation.

---

**Figure 4.** (A) Expression levels from five S. cerevisiae promoters containing wild-type (white) or mutant (black) Ume6 sites after 8 h in sporulation media. (B) Expression levels from four promoters containing wild-type (white) or mutant (black) Ndt80 sites after 8 h in sporulation media. Expression levels from wild-type and mutant promoters in an ume6Δ (C) or ndt80Δ (D) strain after 8 h in sporulation media. The data displayed are the log2 of the mean β-gal expression intensity of five biological replicates. * = P < 0.05; ** = P < 0.01, by Student’s t-test.
Sources (Zhu and Zhang 1999; Cliften et al. 2003; Kellis et al. 2003). To increase the sensitivity of our analysis, we derived a PWM for each of these published motifs (see Methods).

We calculated the probability of each of the 163 motifs being conserved in neutrally evolving sequence. The probability of a TFBS being conserved across the three species is <5% for 134 of 163 motifs, and <1% for 69 of 163 motifs (e.g., Table 2). As expected, the neutral probability was well correlated with information content, which is a function of the size and degeneracy of the TFBS (Fig. 6). For those TFBSs of ≥8 base pairs, 98 of 101 TFBSs had a neutral probability <5%, and 68 of 101 had a neutral probability <1%. Thus, despite their short degenerate nature, most of the TFBSs we tested had a very low probability of being conserved among S. cerevisiae, S. paradoxus, and S. mikatae.

We annotated the S. cerevisiae genome for conserved instances of the 134 TFBSs that have a low probability of conservation. Annotation of 2 Mb of aligned intergenic sequences produced 27,225 predicted TFBSs or 6.5 sites per intergenic sequence. However, most intergenic sequences showed much higher levels of sequence conservation than could be explained by a handful of conserved TFBSs. The ratio of intergenic to synonymous substitution rate implies that 43% of all intergenic sites should rarely be conserved by chance, and we identified 27,225 conserved instances of these sites in 4188 intergenic sequences. Together these conserved TFBSs account for less than half of the functionally constrained intergenic sequences. Mutations in eight of nine predicted Ume6 and Ndt80 binding sites confirm our prediction that functional TFBSs can be identified through sequence conservation alone. However, as we discuss below, not all functional TFBSs may be identified. Additionally, we found that three of five Ndt80 binding sites that are not conserved produce Ndt80-dependent effects on expression when mutated. Whether or not these sites confer any biological differences between species remains an important unanswered question.

The complete results for the 163 motifs are in Supplemental Table 4. GBrowse (Stein et al. 2002) formatted files containing TFBS annotation are available as Supplemental files.

Discussion

We estimated that ∼40% of intergenic sequences in S. cerevisiae are functionally constrained. These sequences may function in transcriptional regulation, translational regulation, or may function as noncoding RNA genes. This estimate is comparable to a previous estimate that 34% of the intergenic sequences are functionally constrained between S. cerevisiae, S. paradoxus, S. mikatae, and S. bayanus (Chin et al. 2005). In the present study we examined the portion of this constraint that can be attributed to conservation of TFBSs. We found that 134 of 163 regulatory motifs should rarely be conserved by chance, and we identified 27,225 conserved instances of these sites in 4188 intergenic sequences. Together these conserved TFBSs account for less than half of the functionally constrained intergenic sequences. Mutations in eight of nine predicted Ume6 and Ndt80 binding sites confirm our prediction that functional TFBSs can be identified through sequence conservation alone. However, as we discuss below, not all functional TFBSs may be identified. Additionally, we found that three of five Ndt80 binding sites that are not conserved produce Ndt80-dependent effects on expression when mutated. Whether or not these sites confer any biological differences between species remains an important unanswered question.
fall just below the threshold for a match to the Ume6 PWM. The SPO12 promoter contains a Ume6 site, but there is no alignment available for its promoter. Of the eight Ndt80 sites, we correctly identified four (SPR3, SPS19, SMK1, and NDT80); SPS4 and SPS1 were not in our alignment data set, and DID2 and CDC10 do not match the Ndt80 PWM, despite matching the middle sporulation element (CRCAAAAY). In total, we correctly identified five of eight known Ndt80 or Ume6 TFBSs for which we have alignments. Thus, our strict motif model and stringent filter for alignment quality make our predictions a conservative estimate of the total Ndt80 and Ume6 regulons.

Annotation of transcription factor binding sites throughout the genome

We found that 133 of 163 known or putative transcription factor binding motifs have a sufficiently low probability of conservation to identify single sites that are under functional constraint. Annotating these constrained sites in the S. cerevisiae genome results in the prediction of 6.5 functional TFBSs per intergenic sequence. However, this annotation relies on a number of assumptions. First, we must assume mutational homogeneity across the genome. Direct estimates of mutation rates are consistent with a uniform mutation rate (Drake 1991), but it is also possible that rare mutational hot-spot or cold-spots may exist (Chuang and Li 2004). In yeast the synonymous substitution rate is nearly uniformly distributed across the genome (Chin et al. 2005). This suggests that mutational heterogeneity does not contribute to the divergence rate among the Saccharomyces genomes.

A second assumption is that substitutions are independent of other positions within a TFBS. Although position independence is widely assumed within TFBSs, this has been proven false for some binding sites (Man and Stormo 2001; Bulyk et al. 2002). When calculating the probability that a TFBS is conserved between species, any dependencies between columns will reduce the number of ways in which a site could be maintained as a TFBS. Therefore, the position-independence assumption makes the probabilities calculated here a conservative estimate of the true probability that a binding site will be conserved in neutral sequence.

A third assumption required for TFBS annotation is that TFBS conservation is due entirely to its functional importance in transcriptional regulation. This assumption may break down under conditions where overlap exists between TFBSs and other functional noncoding sequences. However, only 659 of the conserved TFBSs (2%) were found in 106 noncoding RNA genes (including tRNAs) present in the 4188 intergenic alignments. It is also possible that TFBSs overlap with each other, creating ambiguity about which TFBS is responsible for the functional constraint. The 27,225 significantly conserved TFBSs identified in the three yeast genomes account for ~470 kb of noncoding sequence, of which ~122 kb (27%) is shared between two or more binding sites. Therefore, in many cases we cannot uniquely determine which TFBS is responsible for the functional constraint.

Finally, we assume that the multiple sequence alignments are correct. Simulations have shown that when the number of substitutions is less than one per site, the multiple sequence alignment programs perform quite well (Pollard et al. 2004). We avoided misalignments by choosing closely related yeast species and filtering out any dubious alignments (see Methods). However, some errors due to misalignment may be unavoidable. Even if some misaligned sequences are included in our analysis, this will likely result in missing conserved TFBSs (false negatives), rather than identifying false positives.

The limitations of comparative genomics for identifying functional elements in noncoding DNA

The comparative genomic method applied here and in many other studies relies on the assumption that functional elements will be shared between species. However, it is hypothesized that the majority of biological differences between species are due to changes in gene regulation (Wilson et al. 1974), suggesting that there may be species-specific regulatory signals. The fraction of genes that are differentially regulated between species is not known, but there is evidence to suggest that this fraction may be significant (Dermitzakis and Clark 2002; Moses et al. 2003). Furthermore, many conserved noncoding sequences are only conserved in a subset of taxa (Frazer et al. 2004), suggesting that very closely related species are needed to identify more recently evolved functional sequences (Boffelli et al. 2003). In addition, TFBS turnover (Ludwig et al. 2000; Dermitzakis and Clark 2002; Costas et al. 2003) and misalignment (Pollard et al. 2004) limit the comparative method even when the biology of the two organisms is the same (Ludwig et al. 1998). Our data (Fig. 5) suggest that species-specific regulatory sequences may be common in S. cerevisiae. These sites may be newly evolved Ndt80 sites that serve an S. cerevisiae-specific function, or they may also exist if there are no deleterious consequences of their presence in the promoter sequence. In Drosophila, there is evidence of negative selection against spurious TFBSs (Hahn et al. 2003). Although many Ndt80 binding sites are not conserved, the paucity of unconserved Ume6 binding sites could be explained by negative selection against mutations that create spurious Ume6 binding sites (Fig. 1). The ability to identify TFBS gain, loss, and turnover between species will be crucial to our understanding of the role transcriptional regulation plays in evolution.

The cis-regulatory code

The cis-regulatory code is the set of rules that enables a cell to direct the expression program of each gene based on its cis-regulatory sequences. Identifying functional cis-regulatory sites in the S. cerevisiae genome will be a significant step towards describing this regulatory code. However, our results also suggest that although we may be able to annotate the location of the TFBSs, their effects on gene regulation may often be context-dependent, as shown by the two identical Ume6 sites with opposite effects on expression (Fig. 4). It has been suggested that the presence of Abf1 is required for Ume6 activation (Bowdish and Mitchell 1993), and we found a highly significant concurrence of the conserved Ume6 and Abf1 binding sites (hypergeometric $P < 10^{-18}$). Further experiments will be required to determine in what context (e.g., location from the transcription start site, adjacent regulatory sites, orientation) Ume6 acts as a repressor or an activator. The cis-regulatory code hypothesis also requires that the majority of regulatory elements are not unique to a single promoter. Our estimation that 43% of intergenic sequences are functionally constrained (Table 1) but that only 17% of intergenic sequences match conserved TFBSs suggests that a large fraction of conserved intergenic sequences may serve promoter-specific functions. This is supported by the observation that many experimentally verified regulatory motifs are often flanked by conserved sequences (Fig. 3). A similar pattern is found in
Drosophila, where experimentally verified enhancers are as conserved as their flanking sequences (Berman et al. 2004). The role of promoter-specific sequences in transcriptional regulation needs to be further explored.

**Methods**

**Substitutions in coding and noncoding sequences**

The alignments used for this analysis were downloaded from http://www.broad.mit.edu/annotation/fungi/comp_yeasts/downloads.html (Kellis et al. 2003). The intergenic alignments were filtered to remove any alignment with >10% Ns or 10% missing data in any one species. In the TFBS analysis described here, 4188 alignments were used; 2098 coding sequence alignments were used to generate the synonymous and nonsynonymous rate estimates. The synonymous and nonsynonymous and intergenic substitution rates were estimated using PAML software (Yang 1997). The intergenic substitution rate was estimated using the Hasegawa, Kishino, and Yano substitution model (Hasegawa et al. 1985).

**Motif models**

The initial motif models were obtained from three sources. Cliften et al. (2003) published 65 motifs that were perfectly conserved at least five times between multiple yeast species. Kellis et al. (2003) also identified 71 motifs by their sequence conservation between species. The Kellis et al. (2003) motifs were downloaded from http://www.broad.mit.edu/annotation/fungi/comp_yeasts/motiflist.html. The Saccharomyces cerevisiae Promoter Database (SCP) (Zhu and Zhang 1997) contains 42 motifs identified by various experimental methods. There were a total of 163 motifs after we removed motifs that were exactly identical, as well as shorter motifs that were contained within longer motifs.

A position weight matrix (PWM) was used to describe all motif models. Because the published motif models were only available as consensus sequences, we generated substitution-derived PWMs, based on the observation that the same positions that are degenerate within a species are degenerate between species (Moses et al. 2003). The substitution PWM is based on all instances of the consensus sequence with 0 or 1 difference across the three species (Supplemental Fig. 4). To avoid adding too much noise to the motif model, we tested whether the differences were uniformly distributed across the positions within the motif using a χ² test (d.f. equal to the motif width -- 1). If the differences were uniformly distributed, we used a PWM representing the consensus sequence. The substitution-derived PWMs were found to be very similar to the original motif model except that we were able to refine the frequency of each base at degenerate positions for 91 of the 163 of the motifs. The PWMs used are available in Supplemental File 1.

We used Patser (Hertz and Stormo 1999), with equal base frequencies, to scan all sequences for matches to a PWM. The use of local base frequencies may lead to missing weak sites in regions of biased nucleotide composition (Dermitzakis et al. 2003). For a motif of width w, Patser scored each w-mer as the log-likelihood ratio of observing the sequence under the motif model or the background nucleotide frequencies. If S was greater than the default threshold score, then the w-mer was a match to the motif. The identification of TFBSs and the probability of TFBS conservation were dependent on this cutoff for PWM matches.

As shown in Supplemental Figure S5, the choice of cutoff does not affect the overall percentage of PWMs estimated to be functional.

**Calculating the neutral expectation for TFBS conservation**

We calculated the neutral expectation for TFBS conservation as the probability that a sequence matching the PWM in S. cerevisiae will remain a match to the PWM given a specified amount of evolutionary time. Starting from the highest possible scoring sequence for a particular PWM, we tested all possible sequences to which this initial sequence could evolve. For each descendant sequence, we tested whether or not this sequence was a match to the motif. If it was a match, we calculated the probability of observing this sequence (Fig. 2). The probability of TFBS conservation can be written as:

\[
P(\text{conservation}) = \sum_{x=1}^{w} P(X|x) P(Y|X) P(Z|Y) \delta(\text{match})
\]

where \(x\) is the number of identical bases between sequence \(s\) and the starting TFBS sequence, \(y\) is the number of transitions, and \(z\) is the number of transversions. \(P(X)\) is the probability of a base remaining the same over the given evolutionary distance, \(P(Y)\) is the probability of a transition, and \(P(Z)\) is the probability of a transversion. The probabilities were calculated using Kimura’s two-parameter model (Kimura 1980) with a transition/transversion ratio of two, and the synonymous substitution rate. \(W\) is the width of the motif, and \(\delta(\text{match})\) is a binary function that returns 1 if the sequence was a significant match to the PWM. The probability that a TFBS is conserved across the three-species phylogeny depends on the ancestral state of the three species, which is not known. To avoid this difficulty, we calculated the probability from the sum of the three branch lengths. This simplification results in a conservative estimate of the probability of TFBS conservation.

**Strains, media, and plasmids**

Escherichia coli DH5α was used for all plasmid manipulations. All yeast strains were derivatives of S288C. DBY7468 (α/a, Δura3 EcoRV-Std/ura3-52, his3/his3) was used to measure wild-type expression levels. Expression levels were also measured in ndt80Δ and ura6α strains obtained from the yeast deletion collection (Giaever et al. 2002), present in BY4743 (α/a, his3D1/his3D1, leu2D0/leu2D0, lys2D0/LYS2, MET15/met15D0, ura3D0/ura3D0). YEp357R is a yeast-bacteria shuttle vector maintained in yeast as an episomal plasmid carrying the β-galactosidase gene and the URA3 gene for selection in yeast (Myers et al. 1986). Yeast cultures were grown in complete minimal medium, minus Uracl. Expression was measured in complete minimal medium or sporulation medium (1% potassium acetate).

**Measuring expression from wild-type and mutant promoters**

Promoters containing Ume6 or Ndt80 binding sites were cloned from S. cerevisiae strain S288C into YEp357R to generate a β-GAL fusion construct. The Ume6 or Ndt80 motifs were mutated using the QuikChange Site-Directed Mutagenesis Kit (Stratagene). The alignments of these binding sites and their flanking sequences can be found in Figure 3. Both mutant sequences score well below the Patser threshold and were confirmed by sequencing. Wild-type and mutant promoters were transformed into S288C using lithium acetate (Gietz et al. 1995). Yeast cultures were started at an OD₆₀₀ of 0.2 in complete medium and an OD₆₀₀ of 0.5 in sporulation medium. Expression was measured after 4, 6, 8, and 24 h of growth. The data for growth in complete media...
can be found in Supplemental Figures 2 and 3. The 8-h timepoint is shown in Figures 4 and 5. Data from the other timepoints agree with the 8-h data.

Acknowledgments

We thank Maia Dorsett and Hyun Seok Kim for helpful discussions about the experimental and computational design and procedures, Mark Johnston, Linda Riles, and Jim Dover for providing the deletion strains and plasmids, and Ting Wang, Sudhir Nayak, Sean Eddy, Mark Johnston, and Gary Stormo for helpful comments about the manuscript. S.W.D. is supported by NSF graduate fellowship #DGE-0202737.

References


Web site references

http://www.broad.mit.edu/annotation/fungi/comp_yeasts/downloads.html; yeast intergenic alignments used in this article’s analysis.

http://www.broad.mit.edu/annotation/fungi/comp_yeasts/motiflist.html; cis-regulatory motifs used in this article’s analysis.

Received December 16, 2004; accepted in revised form March 8, 2005.