Evolutionary Forces Act on Promoter Length: Identification of Enriched Cis-Regulatory Elements

Erik Kristiansson,*†‡§ Michael Thorsen,¶ Markus J. Tamás,¶ and Olle Nerman‡§

*Department of Zoology, University of Gothenburg, Göteborg, Sweden; †Department of Neuroscience and Physiology, the Sahlgrenska Academy at the University of Gothenburg, Göteborg, Sweden; ‡Department of Mathematical Statistics, Chalmers University of Technology, Göteborg, Sweden; §Department of Mathematical Statistics, University of Gothenburg, Göteborg, Sweden; ¶Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, Copenhagen, Denmark; and ¶Department of Cell and Molecular Biology/Microbiology, University of Gothenburg, Göteborg, Sweden

Transcription factors govern gene expression by binding to short DNA sequences called cis-regulatory elements. These sequences are typically located in promoters, which are regions of variable length upstream of the open reading frames of genes. Here, we report that promoter length and gene function are related in yeast, fungi, and plants. In particular, the promoters for stress-responsive genes are in general longer than those of other genes. Essential genes have, on the other hand, relatively short promoters. We utilize these findings in a novel method for identifying relevant cis-regulatory elements in a set of coexpressed genes. The method is shown to generate more accurate results and fewer false positives compared with other common procedures. Our results suggest that genes with complex transcriptional regulation tend to have longer promoters than genes responding to few signals. This phenomenon is present in all investigated species, indicating that evolution adjust promoter length according to gene function. Identification of cis-regulatory elements in Saccharomyces cerevisiae can be done with the web service located at http://enricher.zool.gu.se.

Introduction

Interactions between proteins and DNA are central in most aspects of genetic activity including gene transcription and DNA packaging, replication, and repair. Consequently, it is of great importance to further develop the technologies needed to identify and/or predict these interactions. In transcriptional regulation, the protein–DNA interaction consists of the binding of a transcription factor to a specific cis-regulatory element in the promoter of its target genes, thereby regulating the recruitment of the transcriptional machinery. The transcriptional patterns of most genes are therefore dependent on the presence of cis-regulatory elements in their promoters (Tavazoie et al. 1999).

In the recent years, several studies (Hughes et al. 2000; Simon et al. 2001; Nelander et al. 2005; Tsai et al. 2005; Smith et al. 2007) have successfully shown that expression patterns can be used to deduce the transcription factors involved in the transcriptional regulation of genes. This is typically done by identification of coexpressed genes, that is, genes with similar transcriptional profiles, and a consecutive test for enrichments of cis-regulatory elements within the promoters of these genes (Boer et al. 2003; Thorsen et al. 2007). To this end, a number of testing procedures have been suggested, such as the hypergeometric test (Hughes et al. 2000; Sharan et al. 2003; Ettwiller et al. 2005) and tests based on regression models (Bussusmaker et al. 2001; Keles et al. 2004; Copley 2005).

For organisms with a compact genome such as Saccharomyces cerevisiae, it is evident that the length of the intergenic regions varies substantially. This is a potential pitfall that the enrichment studies described hitherto have failed to address. Still, most procedures circumvent the problem by defining the promoter as a fixed number of base pairs upstream of the transcriptional start site, typically 1,000 bp (Gasch et al. 2000; Haverty et al. 2004; Liu and Ringner 2007) but also 800 bp (Daran-Lapujade et al. 2004; Garcia et al. 2004) and 600 bp (Patil et al. 2004) have been used. The median intergenic distance in S. cerevisiae is 455 bp (see Results), but because the variation is large, a fixed promoter length (e.g., 1,000 bp) is a non-optimal solution.

In this paper, we show that functional information is encoded in the promoter length. In particular, we identify several categories of genes with longer or shorter promoters than the rest of the genes in the S. cerevisiae genome. Furthermore, we show that these differences are conserved in several other species which suggests that evolution adapts promoter length according to gene function. Consequently, we propose a novel method to identify relevant cis-regulatory elements in a set of coexpressed genes. The method takes the promoter length into account and can thus perform satisfactorily when there is a difference in promoter length between the set and the remaining genes in the genome. Identification of enriched cis-regulatory elements in S. cerevisiae can be performed with the Enricher web service located at http://enricher.zool.gu.se.

Materials and Methods

Sequence Data and Analysis

The intergenic regions for S. cerevisiae and Schizosaccharomyces pombe were downloaded from the GeneDB database (Hertz-Fowler et al. 2004) (March 2007). For each gene, the 5’ upstream sequence from the start of the open reading frame (ORF) until the end of the previous ORF (on any strand) was extracted. In this study, these sequences are referred to as promoters. Genes with ORFs overlapping other genes such that the intergenic region is nonexistent were removed from the analysis. This resulted in 5,735 promoters in S. cerevisiae (104 genes removed) and 5,484 promoters in S. pombe (379 genes removed). Sequence data for Saccharomyces mikatae, Saccharomyces kudriavzevi, Saccharomyces bayanus, Saccharomyces castellii, and Saccharomyces klyveri, all yeast species closely related to S. cerevisiae (Clifffen et al. 2003), were downloaded from
Promoters for every gene with a homologue in *Saccharomyces cerevisiae* were extracted and cut at the same position as the corresponding promoter in *S. cerevisiae*, resulting in 2,983, 3,606, 4,733, 4,090, and 2,798 promoters for *S. mikatae*, *S. kudriavzevii*, *S. bayanus*, *S. castellii*, and *S. kluveri*, respectively. Promoters for *Ashbya gossypii* were extracted from the *Ashbya* Genome Database (Gattiker et al. 2007) and *S. cerevisiae* homologues were mapped using the table published in Dietrich et al. (2004). Promoters for *Arabidopsis thaliana* were retrieved from the TAIR database (Rhee et al. 2003). The total number of promoters for *A. gossypii* and *A. thaliana* were 4,683 and 27,736, respectively.

All tests of differences in promoter lengths between sets of genes performed in this study were done using the nonparametric Wilcoxon rank sum test. The null hypothesis in the Wilcoxon test assumes that both groups have an equal arbitrary distribution and the test have good power for detecting differences in location. The Wilcoxon test procedure is robust against nonnormal distributions and can therefore handle exaggerated promoter lengths in a proper way. The Holm–Bonferroni procedure was used to correct for multiple testing (Holm 1979).

### Transcription Factor Binding Site Data and Phylogenetic Filtering

In all, 129 known transcription factor binding sites (motifs), corresponding to 92 distinct transcription factors in *S. cerevisiae*, were downloaded from SGD (March, 2007). These motifs were collected from the literature and stored as International Union of Pure Applied Chemistry (IUPAC) codes by the SGD staff. The lengths varied between 5 bp and 20 bp (9 bp in average). In this study, we defined a motif as present within a promoter given that there was at least one exact match. Under phylogenetic filtering (only used in yeast), a motif was defined as present if it was 1) present in the promoter of *S. cerevisiae* and 2) present in at least half of the available corresponding promoters in the closely related yeast species. Note that these filtering rules do not depend on a successful multiple alignment and they correspond to the branch length score suggested by Kheradpour et al. (2007) and Stark et al. (2007) but with uniform branch lengths and a motif movement window equal to the entire promoter. However, the methods for identification of enriched *cis*-regulatory elements described in this paper are applicable to almost any phylogenetic filtering procedure including alignment-based phylogenetic footprinting (Cliften et al. 2003; Kellis et al. 2003).

### Three Tests for Enrichment of Transcription Factor Binding Sites

In this paper, we will use three different procedures to test for enrichment of *cis*-regulatory elements in a set of genes. Assume that there are *N* genes in the genome and that we are interested in testing enrichment of a motif within a subset *A* consisting of *n* genes. For *g* = 1, ..., *N*, let *y*~*g*~ be a binary 0, 1 valued variable indicating whether the motif is present in the promoter of gene *g*. Furthermore, let *x*~*g*~ be another binary variable indicating if gene *g* belongs to the subset *A*. Let *z*~*g*~ = *x*~*g*~ *y*~*g*~, that is, *z*~*g*~ is one if gene *g* is in the subset *A* and has the motif present in the promoter. Finally, let *l*~*g*~ be the length of the promoter for gene *g*.

The hypergeometric test can be seen as drawing a fixed number of balls (genes in *A*) from an urn consisting of red (genes with a motif) and blue (genes without a motif) balls. The observed number of red balls among the drawn balls (genes in *A* with the motif) can then be compared with all other possible draws and the significance calculated. If we let *y*~tot~ = ∑~*g*~ *y*~*g*~, and *z*~tot~ = ∑~*g*~ *z*~*g*~, then the *P* value for the hypergeometric test can be calculated as

\[
P_{\text{hyper}} = \frac{\min(\gamma_{\text{tot}}, n) \binom{N - \gamma_{\text{tot}}}{n - \gamma_{\text{tot}}}}{\binom{N}{n}}.
\]

The hypergeometric test relies on the assumption that all balls are drawn with equal probability, that is, that *cis*-regulatory elements occur with the same probability in all promoters. This assumption is clearly questionable when the promoter length varies between the different genes.

The standard logistic regression model can be formulated as a linear relationship between the log-odds of having the motif present (*y*~*g*~) and the selection of the gene (*x*~*g*~), that is,

\[
\log \frac{\text{Prob}(y_g = 1)}{\text{Prob}(y_g = 0)} = \alpha + \beta x_g.
\]

The coefficients *α* and *β* are estimated from data and an enrichment can be inferred by testing whether *β* is positive. Note that under the null hypothesis, that is, when *β* is zero, the probability of finding the motif is equal for all genes regardless of the length of the promoter.

The method proposed in this paper is an extension of the logistic regression model to a generalized additive model (GAM) (Hastie and Tibshirani 1987, 1990; Wood 2006). In GAMs, nonlinear relations of covariates are modeled nonparametrically by smoothing functions. In our case, an unknown function of the promoter length is added to compensate for any bias in the set of genes of interest. In other words, we assume that the log-odds of the presence of the motif depends both on the selection of the gene and the length of the promoter. Using the same notation as above, the extended model can be formulated as

\[
\log \frac{\text{Prob}(y_g = 1)}{\text{Prob}(y_g = 0)} = \alpha + \beta x_g + f(l_g),
\]

where *f* is an unknown smooth function. The coefficients *α* and *β* and the function *f* are estimated from data. Extremely long promoters, typically coming for regions scarce of ORFs such as the telomeres, were truncated at the 95th percentile resulting in a maximal length of 1,902 bp for *S. cerevisiae*.

All calculations were performed using the statistical language R (R Development Core Team 2008). The mgcv
Details Regarding the Simulation Study

The data for the simulation study were generated as follows. Two sets of artificial promoters were created, one with 500 and the other with 5,000 sequences. For the first group, the promoter lengths were sampled without replacement from the length of the promoter of the common environmental response (CER) genes (Causton et al. 2001). For the second group, the promoter lengths were analogously sampled from the remaining genes in the genome. The difference in median promoter length between the two groups was 107 on average. Given the length, the nucleotide sequence of the promoters was generated by repeatedly sampling from A, C, G, and T with equal probability. In total, 1,000 such sets of promoters were generated (i.e., 1,000 sets where each set has 5,500 promoters in total), and for each set, enrichment of a six-nucleotide motif was inferred using the three methods described above.

Results
Promoter Length in S. cerevisiae

The simplest definition of a gene’s putative promoter region is the DNA sequence located 5’ of the corresponding ORF and stretching to the upstream ORF. Extracting promoter regions for all genes according to this definition revealed the median length of all promoters in the S. cerevisiae genome is 455 bp and that the length of promoters varies greatly (fig. 1).

Promoter Length Diffsers between Functionally Different Categories of Genes

By analyzing different gene categories, we wanted to explore if the promoter lengths contain functional information. In a landmark genomic study in S. cerevisiae, a large set of genes has been identified as differentially expressed in response to a variety of environmental perturbations. This phenomenon was called CER and comprises ~600 genes (Causton et al. 2001). The CER genes are the genes which respond similarly to environmental stimuli, including temperature shock, osmotic stress, oxidative stress, low pH, high pH, and nutrient starvation. Our analyses show that the CER genes have relatively long promoters. The median promoter length for CER genes is 552 bp compared with 445 bp for the promoters of the remaining genes ($P = 4 \times 10^{-5}$) (table 1 and fig. 2). The low $P$ value implies a statistically significant bias toward longer promoters for CER genes compared with other genes in the genome.

Relatively Long Promoters of Stress-Regulated Genes Is a Conserved Phenomenon

To further pursue whether the relationship between promoter length and gene function is an evolutionarily conserved feature, we extended our analysis to include other species. The filamentous fungus A. gossypii has the smallest genome of any free-living eukaryote reported so far. The genome is extremely compact with a median distance between ORFs of only 342 bp. Ashbya gossypii and S. cerevisiae diverged more than 100 Ma and their genomes differ substantially in GC content. Still, 95% of the protein-coding sequences of A. gossypii have homologues in the S. cerevisiae genome with the vast majority at syntenic locations (Dietrich et al. 2004). We could therefore find homologs for 405 of the 588 CER genes in the A. gossypii genome. These putative CER genes have a median promoter length of 407 bp, which is significantly longer than the median promoter length of remaining genes of 337 bp ($P = 3 \times 10^{-5}$) (table 1).

The evolutionary divergence between the fission yeast S. pombe and the budding yeast S. cerevisiae is estimated to 1.14 billion years (Hedges 2002), almost as great as the distance between S. cerevisiae and man. Therefore, an analysis of the promoters in the S. pombe genome could be a better indication for any potential evolutionary constraints on
promoter length. We extracted the promoters of the S. pombe genome similarly to our S. cerevisiae approach. The median length of the promoter regions in the S. pombe genome is 829 bp, and the length of promoters varies greatly (supplementary fig. 1, Supplementary Material online). Analyses of stress-associated genes in S. pombe revealed that many genes display analogous transcriptional profiles in response to oxidative stress, cadmium stress, temperature shock, osmotic stress, and a DNA-damaging agent. Here, the authors named it the core environmental stress response (CESR) (Chen et al. 2003). A conservative definition (regulated at least 2-fold in four of the five stress conditions) of the CESR comprises 237 genes, whereas a more loose definition (regulated 2-fold in at least one stress condition) resulted in 704 genes. With the loose definition, the CESR genes have a median promoter length of 863 bp compared with the median of the remaining promoters of 826 bp ($P = 4 \times 10^{-5}$). However, analyzing the more conservatively defined CESR genes gives a more striking result. Here, the median length of the regulated genes is 1,243 bp, whereas the median length of the unregulated genes is 816 bp ($P = 6 \times 10^{-13}$) (supplementary fig. 2, Supplementary Material online).

Next, we analyzed promoter lengths in the genome of the plant A. thaliana. Identification of genes transcriptionally upregulated in response to environmental stress treatments (heat, cold, drought, salt, high osmolarity, UV-B light, and wounding) has been reported (Kilian et al. 2007). The median promoter length for these A. thaliana environmental stress-responsive genes was 1,672 bp compared

### Table 1

<table>
<thead>
<tr>
<th>Gene Subset</th>
<th>Number of Genes</th>
<th>Median Length</th>
<th>Number of Remaining Genes</th>
<th>Median Length Remaining Genes</th>
<th>Adjusted P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Saccharomyces cerevisiae</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Causton CER genes (microarray)</td>
<td>588</td>
<td>552</td>
<td>5,147</td>
<td>445</td>
<td>$4 \times 10^{-9}$</td>
</tr>
<tr>
<td>Thorsen upregulated arsenite genes (microarray)</td>
<td>317</td>
<td>560</td>
<td>5,418</td>
<td>450</td>
<td>$1 \times 10^{-5}$</td>
</tr>
<tr>
<td>Thorsen downregulated</td>
<td>398</td>
<td>561</td>
<td>5,337</td>
<td>448</td>
<td>$2 \times 10^{-5}$</td>
</tr>
<tr>
<td>Essential genes</td>
<td>1,033</td>
<td>385</td>
<td>4,702</td>
<td>468</td>
<td>$7 \times 10^{-7}$</td>
</tr>
<tr>
<td>Spellman cell cycle genes</td>
<td>726</td>
<td>563</td>
<td>5,009</td>
<td>443</td>
<td>$4 \times 10^{-12}$</td>
</tr>
<tr>
<td>GO: Cell wall</td>
<td>69</td>
<td>899</td>
<td>5,666</td>
<td>453</td>
<td>$2 \times 10^{-9}$</td>
</tr>
<tr>
<td>GO: Transporter activity</td>
<td>361</td>
<td>583</td>
<td>5,374</td>
<td>449</td>
<td>$3 \times 10^{-7}$</td>
</tr>
<tr>
<td>GO: RNA metabolic process</td>
<td>262</td>
<td>340</td>
<td>5,473</td>
<td>461</td>
<td>$4 \times 10^{-7}$</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chen CESR genes (microarray)</td>
<td>704</td>
<td>863</td>
<td>4,780</td>
<td>826</td>
<td>$9 \times 10^{-1}$</td>
</tr>
<tr>
<td>Chen CESR conservative genes</td>
<td>237</td>
<td>1,243</td>
<td>5,247</td>
<td>816</td>
<td>$6 \times 10^{-13}$</td>
</tr>
<tr>
<td>(microarray)(microarray)(microarray)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linder med-12 mediator genes</td>
<td>173</td>
<td>1,287</td>
<td>5,311</td>
<td>819</td>
<td>$5 \times 10^{-12}$</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kilian stress genes (microarray)</td>
<td>1,152</td>
<td>1,672</td>
<td>26,584</td>
<td>1,113</td>
<td>$&lt;10^{-16}$</td>
</tr>
</tbody>
</table>

*a For growth in rich medium.

b Homologues based on the map in Dietrich et al. (2004).

c Genes from the environmental stress microarray data in Kilian et al. (2007) with an log fold-change more than 5 in at least one condition.
with the 1,113 bp of the promoters of the remaining genes (P < 10^{-16}) (table 1). Taken together, these findings suggest that promoter length and gene function are correlated.

Promoter Length Varies for Many Gene Categories

Testing whether other biologically meaningful gene sets not related to stress responses have divergent length of promoter regions, we found that this is in fact a common phenomenon. The transcription of many genes is known to oscillate in synchronicity with the cell cycle. Comprehensive identification of cell cycle–regulated genes in *S. cerevisiae* of Spellman et al. (1998) yielded a list of 726 genes. The promoters of these genes have a median length of 564 bp and are thus significantly longer than the median of 443 bp of the remaining promoters in the genome (table 1).

The mediator is an evolutionarily conserved multisubunit protein complex which works as an interface between gene-specific transcription factors and RNA polymerase II–dependent genes (Kornberg 1996). In *S. pombe*, inactivation of individual mediator subunits affects the transcriptional level of different subsets of mediator-controlled genes (Linder et al. 2008). Using the data from Linder et al., we found that 173 genes are affected in transcriptional level by deletion of *MED12* (using a 2-fold change cutoff). Investigating the promoter length of these genes differentially expressed in cells mutated in mediator subunits revealed that these genes have long promoters (P = 5 × 10^{-12}) (table 1). Analyses of promoter length of the genes affected by other mutations in mediator components revealed similar tendencies (data not shown).

The examples given so far have been gene sets derived from microarray analyses, so next, we examined the promoter lengths of other types of gene sets. Interestingly, the promoters of several categories of genes were found to be significantly longer or shorter compared with the rest of the genome (table 1). According to the SGD (Cherry et al. 1997), 1,033 genes are essential for growth in rich medium. Our analysis revealed that these genes have very short promoters. The median promoter length for the essential genes is 385 bp, whereas the median promoter length for nonessential genes is 468 bp (P = 7 × 10^{-7}). When we look at *A. gossypii* homologues of essential *S. cerevisiae* genes, we find also their promoters to be much shorter than the promoters of nonessential genes (P = 3 × 10^{-7}). Investigating whether functional categories of genes differ in their promoter lengths, for example, genes annotated similarly in the Yeast GO Slim terms (Ashburner et al. 2000), we found that most functional categories do not exhibit any bias. However, a few categories such as cell wall (median length 899 bp, P = 2 × 10^{-9}), transporter activity (median length 583 bp, P = 3 × 10^{-7}), and RNA metabolic process (median length 340 bp, P = 4 × 10^{-7}) do differ significantly in their promoter lengths. Lists of all GO Slim terms and their corresponding promoter lengths are available as supplementary information (supplementary table 1, Supplementary Material online).

**Correlation between Promoter Length and Cis-Regulatory Elements**

The examples presented above warrant caution when trying to assess enrichment of cis-regulatory elements in promoter sequences. Naturally, the likelihood of a random occurrence of any motif increases with the promoter length as illustrated in figure 3(A). In this figure, the number of different cis-regulatory elements and the promoter length are shown for all genes in the genome of *S. cerevisiae*. Nonetheless, the trend is still present when spurious hits are removed by phylogenetic filtering (fig. 3B), suggesting that this might be a biologically relevant phenomenon.

In a reversed approach, we analyzed the transcriptional profiles of the genes in the CER set and divided all genes into classes depending on the number of stress conditions to which they respond. Plotting the median promoter length of
each group against the number of stress conditions, a significant positive trend is observed ($P < 10^{-5}$, fig. 4). The same was done for the A. thaliana promoters and again we found a positive trend between the promoter length and the number of different conditions which induce transcription of the gene (supplementary fig. 5, Supplementary Material online). This analysis shows that the length of the promoters correlates positively with the number of signals that is integrated onto the promoters.

Assessing Enrichment of Cis-Regulatory Elements under Promoter Length Bias

The systematic variation in promoter length constitutes a problem when testing for enrichment of known cis-regulatory elements in a set of genes, for example, coexpressed genes from a microarray experiment. Traditional methods, such as the hypergeometric test and the logistic regression model, are based on the questionable assumption that the elements can occur with equal probability in each promoter. Because genes with long promoters will, on average, have more random occurrences than genes with short promoters, these methods will use a false null hypothesis and as a consequence the $P$ values become biased. We therefore suggest a novel method to identify enriched cis-regulatory elements in the promoters of a set of genes. The method includes the promoter lengths as a covariate and can thus handle discrepancies in promoter length in a proper manner. The model is based on the GAM framework (Hastie and Tibshirani 1990; Wood 2006), and in contrast to the logistic regression model, it can estimate how the promoter lengths will affect the probability of the occurrence of a cis-regulatory element. Full mathematical details can be found in Materials and Methods.

Promoter Length Bias Can Result in False Positives

Simulated data were used to compare the proposed model to the hypergeometric test and the logistic regression model. Two sets of artificial promoters were generated, one containing 500 sequences with lengths sampled from the set of CER genes and one with 5,000 sequences with lengths sampled from the remaining genome of S. cerevisiae. Tests for enrichment of a six-letter-long element were performed within the smaller group using the three different test procedures. Histograms of the resulting $P$ values for all three methods are shown in figure 5. Because the promoter sequences were completely random, no significant enrichment should exist and the $P$ values should therefore be uniformly distributed. This is clearly not the case for the hypergeometric test and logistic regression model, which both have several low $P$ values, indicating a substantial amount of false positives. The proposed method has, on the other hand, no such tendencies which suggest that it produces fewer false positives when there is a difference in promoter length.

Transcription Factors in Arsenic Stress

To evaluate if the proposed method has any impact on real biological data, we assessed enrichments of cis-regulatory elements in a microarray data set on transcriptional changes in response to arsenite exposure (Thorsen et al. 2007). Based on this experiment, putatively up- and downregulated genes with an absolute log$_2$ fold-change greater than 2 where chosen for enrichment analysis. The promoter lengths of these genes were found to be significantly longer than the promoters in the rest of the genome (table 1). Accordingly, the model proposed in this paper is suitable for the analysis of this data.

All experimentally verified cis-regulatory elements in SGD (Cherry et al. 1997) were individually tested for enrichment using the proposed model. Phylogenetic filtering was applied to remove as many spurious elements as possible (see Materials and Methods). Examples of significant elements can be seen in table 2 together with $P$ values from the hypergeometric test and logistic regression. In general, the $P$ values for highly significant elements seem to be lower for the regression-based methods than for the hypergeometric test, where the proposed method, as expected, is the most conservative. The last two rows in table 2 show examples of elements that are, according to the proposed models, false positives. These elements have low $P$ values for both the hypergeometric test and the logistic regression model but get substantially higher $P$ values when the promoter length is included. The difference between the results shows the large impact of including the promoter lengths when testing for enrichment of cis-regulatory elements. Complete lists with $P$ values for all cis-regulatory elements in yeast can be found in supplementary tables 2 and 3 (Supplementary Material online).

Discussion

Using the S. cerevisiae genome as a model, we analyzed the length of promoter regions of genes in different data sets and found a striking relationship between long promoters and responsiveness to a variety of stress conditions. Complementary to this, we analyzed analogous data...
sets of promoters in *S. pombe* and *A. thaliana* and found the same relationship. Thus, longer promoters of stress-inducible genes seem to be a conserved phenomenon. The simplest explanation may be that relatively long promoters allow for integration of a large number of regulatory inputs onto the promoter region of these genes and, conversely, that other promoters may have been shortened to reduce genome size.

The integration of many signals into a promoter requires a number of elements. If several signaling pathways target the same promoter, binding sites must exist for all relevant transcription factors. Additionally, the regulation can be combinatorial and thus a number of auxiliary factors may also bind. In these cases, the promoters must be long or steric hindrances may prevent the binding of all the necessary factors. We speculate that the level of complexity of the regulation of a gene is related to the function of the gene product. Hence, promoters of certain classes of genes may have evolved to be longer due to complex regulation, whereas other classes of genes may have evolved shorter promoters to reduce genome size. How evolution can result in increased promoter length is intuitive as it allows for the integration of a large number of regulatory inputs and thus a more complex regulation of gene expression. That evolution will sometimes favor a decrease in the length of certain promoters is less obvious as the effect on overall genome size may be relatively minor. The shortening of promoter length can however be an important, albeit small, factor in genome evolution because a smaller genome is believed to be an advantage as replication is faster, allowing for faster proliferation.

Evolution may not have acted to reduce the intergenic stretches of all genomes. Therefore, a relationship between promoter length and regulatory complexity might not always exist. We have argued here that utilizing the information about the length of the individual promoters improves the validity of the *P* values generated when estimating enrichments of *cis*-regulatory elements in *S. cerevisiae*, *S. pombe*, *A. thaliana*, and *C. elegans* data sets. Similar analyses of mammalian promoters may be difficult, as they in general are much longer and less clearly defined. However, the annotation of mammalian genomes is improving (Birney et al. 2007) and several new resources for extracting promoter sequences from mammalian genomes exist (Wakaguri et al. 2008). One interesting study on the organization of the human genome clearly demonstrates that both the introns and the intergenic space around housekeeping genes are shorter than corresponding sequences for tissue-specific genes. Similar trends are also reported for the fly *Drosophila melanogaster* and the worm *Caenorhabditis elegans*. The author speculated that the tissue-specific genes have a more composite transcriptional

<table>
<thead>
<tr>
<th>Table 2</th>
<th>Enriched <em>cis</em>-Regulatory Elements in the Arsenite Microarray Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcription Factor</td>
<td>Binding Site</td>
</tr>
<tr>
<td>RPN4</td>
<td>GGTGCAAA</td>
</tr>
<tr>
<td>YAP2</td>
<td>MITASTMARC</td>
</tr>
<tr>
<td>YAP7</td>
<td>MTKA</td>
</tr>
<tr>
<td>YAP1</td>
<td>TTAGMAGC</td>
</tr>
<tr>
<td>MSN2/4</td>
<td>AAGGGG</td>
</tr>
<tr>
<td>YAP2/3/4/5</td>
<td>TTACTAA</td>
</tr>
<tr>
<td>YAP1</td>
<td>TTASTMA</td>
</tr>
<tr>
<td>MET31/32</td>
<td>AAACGG</td>
</tr>
<tr>
<td>SWI5</td>
<td>KGCCTGGR</td>
</tr>
<tr>
<td>YOX1</td>
<td>YAATTA</td>
</tr>
</tbody>
</table>
regulation and thus require longer sequences to facilitate a more dynamic chromatin structure (Vinogradov 2004). These findings corroborate our conclusions and indicate that promoter length is an important factor when analyzing genomes of higher eukaryotes.

It is far from trivial to strictly define a promoter region. Due to alternative transcription start sites (David et al. 2006) and regions with other, today unknown, function, it is hard to tell exactly what parts of the intergenic regions are included in regulation of transcription. In this study, we have used the entire region between ORFs as a putative promoter, including the untranslated regions (UTRs). The differences in promoter lengths are still present if the UTRs (David et al. 2006) are removed (supplementary table 4, Supplementary Material online), indicating that the UTR regions play a minor role in the observed variation in promoter length.

Assessing enrichments of de novo elements is a process where the rate of false positives has been particularly bothersome. A recent report (Eden et al. 2007) identified a number of problems which need to be addressed to improve these methods. Among them were 1) the lack of rigorous models and of an exact $P$ value measuring motif enrichment and 2) the tendency, in many of the existing methods, to report presumably significant elements even when applied to randomly generated data. We believe that our model is a novel approach to (1) and a clear improvement of (2).

Eukaryotic genomes are packaged into nucleosome particles that prevent the DNA from interacting with other DNA-binding proteins. Experimental evidence shows that DNA packaging can control accessibility of specific sequences by blocking access to irrelevant nonfunctional sites (Sekinger et al. 2005; Yuan et al. 2005). A future direction would therefore be to incorporate chromatin structure prediction into the enrichment analysis. Additional information could be extracted by predicting acetylation and methylation of nucleosomes to identify sequences with regulatory activity (Heintzman et al. 2007). These are important areas of research, and further development of these tools is needed to improve their accuracy. Combining robust prediction algorithms for nucleosome organization with phylogenetically based comparative genomics holds great promises for vastly enhanced prediction of true cis-regulatory elements.

We have presented evidence that promoter lengths include functional information and several categories of genes have been shown to have promoters longer or shorter than promoters of the other genes in the genome. This was seen for sets of genes based on a number of different functional categories in S. cerevisiae, A. gossypii, S. pombe, and A. thaliana, and thus, evolutionary forces act on promoter length. Consequently, we have developed a method for assessing enrichments of cis-regulatory elements in sets of promoters with varying lengths. To our knowledge, this is the first model which utilizes information encoded in the promoter lengths to find overrepresentations of transcription factor binding sites. Simulations show that the proposed model performs adequately, in contrast to the hypergeometric test and logistic regression, which both can lead to a high number of false positives.

**Supplementary Material**

Supplementary figures 1–7 and tables 1–4 are available online at Molecular Biology and Evolution (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

This research was financed by the Swedish National Research School in Genomics and Bioinformatics, the Gothenburg Stochastic Centre, and the Gothenburg Mathematical Modeling Centre. We would like to thank Alexandra Jauhiainen for extensive comments and suggestions on the manuscript and Jonas Warringer and Olga Kourtchenko for discussions.

**Literature Cited**


Douglas Crawford, Associate Editor

Accepted February 24, 2009